

The Role of Evidence Based Prosopography in the Digital Study of Past Human Lives

By

Kelvin Beer-Jones

A thesis submitted to the University of Birmingham for the degree of

Doctor of Philosophy

School of Philosophy, Theology and Religion

College of Arts and Law

University of Birmingham

May 2025

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

Abstract

The Semantic Web presents an opportunity to study Past Human Lives (PHL) and their relationships. Information about PHL resides largely in the physical records in the archives and in digital representations on genealogical platforms. Information about PHL is not Born Digital, so to study PHL digitally, physical information in its original form must first be digitally represented. This is challenging because information about PHL is mostly unstructured, messy and incomplete, and therefore largely unsuitable for digital study as data.

This thesis addresses the problematic relationship between ‘information’ (physical) and ‘data’ (digital) and the lack of comprehensive infrastructure provision for researchers into PHL in digital form. This concern is addressed by showing how extending digital infrastructure to embrace Evidence Based Prosopography (EBP), organised in a National Authority Indexing system and based on Unique Identifiers (NAI-UIDs), ensures that information about PHL is represented digitally in a structured and useful format, thus enabling researchers to take full advantage of the Semantic Web. Issues such as fixity, affixedness and provenance tracking are addressed. The EBP system is based on linking existing data matching practices in genealogy, and existing catalogue and finding aids in academia, and it utilises UID structures common to both. Current Digital Humanities (DH) infrastructure is shown to be sufficiently mature to now extend its provision by adopting the EBP and NAI system.

This study is informed by a digital research project that models relationships between 3000 activists from 1830 to 1870, with 600 Quakers among them. This included design and build

of a Human Data Digital Toolkit (HDDT) to organise and manage an EBP dataset combined from several archival sources in a variety of data formats. This 'Independent Researcher' project was compared with five other contemporaneous (larger and well resourced) research affordances taken from different countries, to show how the Independent Researcher model is not impaired when compared to much larger projects, and that the Independent Researchers needs bridge the gap between academia and genealogy.

The study concludes by demonstrating how the extension of infrastructure provision to include the EBP system will provide significant improvement to infrastructure support for researchers, from large well-resourced down to Independent Researcher projects, across academia and genealogy, and across society. The EBP system addresses key issues in data management, especially the three essential relationships between information and data – fixity, affixedness and provenance. It also provides a structured infrastructure which supports data interoperability and sustainability. The practical next steps necessary to develop the EBP system are set out.

Acknowledgements

I thank my supervisors Professor Ben Pink Dandelion and Professor Hugh Houghton, both of University of Birmingham; Professor Zoë Laidlaw, University of Melbourne who inspired the Project Seven data study into the 3000 persons (and the 600 Quakers amongst them) whose concern for the plight of aborigines led in part to the institutionalisation of anthropology in Britain (1830-1870); Professor Charles van den Heuvel, University of Amsterdam, who provided guidance in the Information Science aspects of the study.

Benjamin Beck, Secretary of the Quaker Family History Society, identified 600 Quakers in the Project Seven database and found the family relationships between them. He also provided advice on working with person names in genealogical research. Mike Allaway, Research Software Engineer, University of Birmingham provided dedicated training and data management advice throughout Project Seven. Jonathan Blaney, Institute of Historical Research, University of London and Martin Steer, Technical Lead, Digital Humanities School of Advanced Study, University of London provided steering and technical advice for Project Seven.

Lastly I thank Dr Sally Osborn of EditExpert who kindly proofread the text.

Table of Contents

Chapter 1 The digital study of Past Human Lives	1
1.1 The Digital Humanities background.....	1
1.2 Definitions of key terms.....	8
1.3 Project Seven (P7) Introduction.....	15
1.4 Researcher concerns identified in P7	17
1.5 Evidence Based Prosopography.....	19
1.6 EBP and the National Authority Index system.....	22
1.7 Thesis objectives	25
1.8 Chapter outlines.....	31
Chapter 2 Evidence Based Prosopography (EBP)	35
2.1 Introduction	35
2.2 The DH context of EBP	36
2.3 Prosopographical Information.....	41
2.4 Digital Evidence Based Prosopography.....	44
2.5 An example of EBP from Project Seven	47
2.6 The role of EBP in genealogy	51
2.7 The proposed National Authority Index system.....	60
2.8 Chapter Summary	77

Chapter 3 EBP in the Digital Humanities.....	79
3.1 The roots of DH are deep in the past.....	83
3.2 The history of DH is characteristically technological	85
3.3 Defining Digital History	92
3.4 Research infrastructure provision	106
3.5 Research technologies, the database	126
3.6 Historians / data scientists.....	140
3.7 Research Software Engineers (RSEs)	144
3.8 Chapter summary	153
Chapter 4 EBP in Galleries, Libraries, Archives, Museums and Special collections (GLAMS).....	156
4.1 Classification of EBPI Sources in GLAMS.....	161
4.2 Leading institutions in GLAMS digital development.....	173
4.3 Metadata and data control in GLAMS	181
4.4 Standards and Conceptual Models in GLAMS	191
4.5 Chapter summary	216
Chapter 5 EBP in recent research projects	218
5.1 Introduction	218
5.2 Social Networks and Archival Context (SNAC).....	223
5.3 The Cambridge Group for the History of Population and Social Structure (The Group) and the I-CeM project.....	241

5.4 Traces Through Time.....	252
5.5 ResearchSpace	257
5.6 The Golden Agents.....	272
5.7 Research projects review	290
5.8 Durability lessons from the five research offerings studied.....	305
Chapter 6 EBP in Project Seven (P7)	309
6.1 Project Seven objectives	312
6.2 Project Seven background	314
6.3 Thomas Hodgkin MD (1789-1866).....	319
6.4 Quaker roles in the Quaker Led Group (QLG).....	321
6.5 Issues in defining the QLG members	326
6.6 The Centres for the Emergence of the Discipline of Anthropology in Britain (CEDA).329	
6.7 The Quaker Committee on the Aborigines.....	330
6.8 The Select Committee on the Aborigines 1834 - 1837	331
6.9 The Aborigines Protection Society.....	331
6.10 The Ethnological Society of London.....	332
6.11 The QLG and the emergence of the discipline of anthropology	333
6.12 Quaker families	335
6.13 The HDDT design challenges.....	335
6.14 Complex prosopographical networks	337

6.15 P7 Topology.....	339
6.16 Designing the HDDT	341
6.17 Building the HDDT	343
6.18 Data collection, management and cleaning	348
6.19 Data analysis and visualisation	368
6.20 Case Study 1 The Centres for the Emergence of the Discipline of Anthropology (CEDA)	382
6.21 Case Study 2 The CEDA Quakers and their relationships	395
6.22 Case Study 3: Thomas Hodgkin’s MD networks	409
6.23 Project Seven review	422
6.24 Project Seven conclusion	429
Chapter 7 Conclusion	430
7.1 Reflection	430
7.2 Returning to the research questions	438
7.3 A data science perspective on the study of Past Human Lives.....	442
7.4 Recommendations	445
7.5 Impact Statement	448
Appendix 1 Thomas Hodgkin MD forms the Aborigines Protection Society	450
Appendix 2 Thomas Hodgkin MD’s political network	453
Appendix 3 Archival research into the Quaker Committees on the Aborigines	466

1	Introduction	466
2	Table of manuscripts.....	469
3	The Minute Books of London Yearly Meeting and the Minute Books of Meeting for Sufferings (1831 – 1846)	477
4	Description of the Quaker archives	469
5	1834 - 1835 The Epistle from North Carolina	496
6	1837 The Select Committee Report, the establishment of the Committee on the Aborigines and the Aborigines Protection Society.	498
7	1838 Upper Canada	501
8	1839 Thomas Hodgkin - The Committee on the Aborigines, and the memorial to Lord John Russell.....	503
9	1842 The enlargement of the Committee on the Aborigines	506
10	1845 - 1846 The end of the Committee on the Aborigines	507
11	Lists of names of Quaker committee members.....	510
12	Conclusion	511
13	Publications of the Quaker Committee on the Aborigines	514
Appendix 4	Records in Context. Definition of terms – Person.....	516
Appendix 5	Archival Records Standards - ISAAR – CPF Second Edition 2004	519
Appendix 6	Describing Archives: A Content Standard (DACS)	537
Appendix 7	The Project Seven Report.....	547

List of References.....	770
-------------------------	-----

Figures

Figure 1.1 Example of a National, Archival and Research Authority Index system based on BMD certificates.....	25
Figure 2.1 The story of smart and big data (Schöch 2013, 11)	39
Figure 2.2 The NAI-UID system example	50
Figure 2.3 Finding EBP in the HSN Project (Kennard, Kent, and Barrett 2011, 49)	58
Figure 2.4 The author's birth record.....	62
Figure 2.5 Free UK Genealogy CIO front page	63
Figure 2.6 The proposed National Authority Index	64
Figure 2.7 The proposed Archival Authority Index.....	66
Figure 2.8 The research dataset component of the proposed National Authority Index system	68
Figure 2.9 The complete NAI-UID system of indexes and UIs – the system works 'top down' and 'bottom up'	70
Figure 2.10 The NAI-UID system: the past world of 'information' becomes 'data' in the digital world	72
Figure 2.11 Example of a Resource Description Framework Schema (RDFS) (Riley 2017, 12).....	74
Figure 2.12 The ARK Alliance: 20 years, 850 institutions, 8.2 billion persistent identifiers (https://www.slideshare.net/jakkbl/the-ark-alliance-20-years-850-institutions-82-billion-persistent-identifiers-20211022 , Accessed 12 September 2023)	75
Figure 2.13 A simple XML example (Riley 2017, 9).....	77

Figure 3.1 Digital technology and social change: the digital transformation of society from a historical perspective (Hilbert 2020, 192)	85
Figure 3.2 The early history of digital humanities: an analysis of computers and the humanities (1966–2004) and literary and linguistic computing (1986–2004) (Sula and Hill 2019, 193)	88
Figure 3.3 Digital humanities methodological commons (after McCarty & Short (2002), (Anderson, Blanke, and Dunn 2010, 3783).	89
Figure 3.4 ‘A rough two-dimensional representation of variations in historical approaches looks like this’ (Tilly 1990, 695).....	95
Figure 3.5 The monographs by Ginzburg, Thompson, Wrigley and Schofield, and Zunz.....	96
Figure 3.6 JISC scope and services (https://www.jisc.ac.uk/about-us , Accessed 1 June 2023)	111
Figure 3.7 University of Birmingham RDM online infrastructure (https://intranet.birmingham.ac.uk/as/libraryservices/library/research/rdm/index.aspx , Accessed 15 October 2023).	117
Figure 3.8 ‘Understanding the information requirements of arts and humanities scholarship’ (Benardou 2010, 23).	121
Figure 3.9 A model of interdependent concepts of digital tool criticism (Koolen, Van Gorp, and Van Ossenbruggen 2019, 373).....	128
Figure 3.10 ‘The spinning plates of family, friendship and acquaintance that circled the time revolution. The result was a unification of knowledge where natural and human history	

were drawn together at different timescales.’ Original artwork Kaylea Raczkowski-Wood (Gamble 2021, 19)	130
Figure 3.11 The author’s schema of his CEDA digital database.....	131
Figure 3.12 An example of a seventeenth-century database: The Table of Casualties, England in the seventeenth century (Bowker and Star 2000, 22).....	133
Figure 3.13 Image of the room of the Enlightenment, British Museum (https://www.britishmuseum.org/collection/galleries/enlightenment , Accessed 1 November 2023)	134
Figure 3.14 Langren’s 1644 graph of determinations of the distance, in longitude, from Toledo to Rome. Taken from Tufte 1997, 15; (Friendly 2008a, 4).	138
Figure 3.15 Recommended dual-project model for collaboration (Breure, Doorn, and Boonstra 2006, 98).....	140
Figure 3.16 The Breure, Doorn and Boonstra model modified by Kelvin Beer-Jones.....	141
Figure 3.17 Researchers by discipline (Olivier et al. 2019, 23)	145
Figure 3.18 Discipline of survey respondents (Olivier et al. 2019, 27)	146
Figure 3.19 Types of research data used by arts and humanities researchers (Olivier et al. 2019, 30)	147
Figure 3.20 Frequency of data sharing – all respondents (Olivier et al. 2019, 36).....	148
Figure 3.21 Frequency of data sharing by humanities/DH researchers (Olivier et al. 2019, 36)	149

Figure 3.22 Aspects of data sharing by humanities/DH researchers (Olivier et al. 2019, 38)	150
Figure 3.23 Data storage locations – all respondents (Olivier et al. 2019, 39)	151
Figure 4.1 Graphic representation of the history of the development of norms and models of archival description since the end of the 1980s (Llanes-Padrón and Pastor-Sánchez 2017, 391)	178
Figure 4.2 How important and useful are standards? (International Council on Archives 2021, 5)	179
Figure 4.3 How far are these standards implemented in your archive? (International Council on Archives 2021, 6)	180
Figure 4.4 What other standards do you use to describe your archives? (International Council on Archives 2021, 6)	181
Figure 4.5 Model of the levels of arrangement of a fonds (International Council on Archives 1999, 37)	189
Figure 4.6 Relationship between descriptive and authoritative records (International Council on Archives 1999, 38)	190
Figure 4.7 Left: Representation of data in a hierarchical structure like XML or other markup-language. Right: Representation of data in triples that results in a graph structure (International Council on Archives 2023, 6)	207
Figure 4.8 Overview of the RiC Conceptual Model – note RiC E07 Agent class (top right) (International Council on Archives 2023, 18)	208

Figure 4.9 ‘The primary role of the CIDOC CRM is to serve as a basis for mediation of cultural heritage information and thereby provide the semantic “glue” needed to transform today’s disparate, localised information sources into a coherent and valuable global resource’ (Oldman 2014)	210
Figure 4.10 CIDOC CRM encoding example (Winkelmann seeing Laocoön) (Bekiari 2021, 36)	212
Figure 4.11 Symbolic representation of ‘Winkelmann seeing Laocoon’ as an evolution in space and time (Bekiari 2021, 37)	213
Figure 4.12 Authorities come from ‘Background knowledge’ in the CIDOC CRM (Stead 2008)	214
Figure 5.1 SNAC is based on the EAC-CFP Schema (https://portal.snaccooperative.org/node/371 , accessed 19 August 2024)	227
Figure 5.2 SNAC History Research Tool (https://snac-web.iath.virginia.edu , accessed 10 August 2024)	228
Figure 5.3 SNAC advanced search (https://snaccooperative.org/search#results , accessed 19 August 2024)	229
Figure 5.4 SNAC main return for person enquiry Thomas Hodgkin (1798–1866) (https://snaccooperative.org/view/61687138 , accessed 07 August 2024)	230
Figure 5.5 SNAC Person Record for Thomas Hodgkin (1798–1866)	231
Figure 5.6 SNAC data sources for person name enquiry Thomas Hodgkin (1798–1866) (https://snaccooperative.org/view/61687138 , accessed 07 August 2024)	232

Figure 5.7 SNAC – enquiry into the name ‘Thomas Hodgkin’, degree 1 (https://snaccooperative.org/visualize/connection_graph/61687138/9024591 , accessed 7 August 2024)	234
Figure 5.8 SNAC – enquiry into the name ‘Thomas Hodgkin MD’, degree 2 (https://snaccooperative.org/visualize/connection_graph/61687138/9024591 , accessed 7 August 2024)	235
Figure 5.9 SNAC – enquiry into the name ‘Thomas Hodgkin’, degree 3 (https://snaccooperative.org/visualize/connection_graph/61687138/9024591 , accessed 7 August 2024)	235
Figure 5.10 University of Birmingham (a hosting of Exlibris) finding aid search: Thomas Hodgkin MD (https://birmingham-primo.hosted.exlibrisgroup.com/primo-explore/search?query=any,contains,Thomas%20Hodgkin%20MD&tab=local&search_scope=CSCOP_44BIR_DEEP&vid=44BIR_VU1&offset=0 , accessed 10 August 2024).	239
Figure 5.11 1851 Household Schedule England and Wales (English) – reverse. Instructions for enumerators (Higgs et al. 2021, 23)	242
Figure 5.12 1851 Enumeration Book – reverse (Higgs et al. 2021, 34)	242
Figure 5.13 Traces through Time project record for Alfred Frederick Minall showing TTT links to ‘other possible matches’ (https://discovery.nationalarchives.gov.uk/details/r/D7695214 , accessed 12 October 2024).....	253
Figure 5.14 Graph showing links across records of Alfred Minall (https://blog.nationalarchives.gov.uk/making-connections-tracing-people-collection , accessed 12 October 2024).....	254

Figure 5.15 ResearchSpace platform architecture showing metaphactory as the comprehensive platform (Oldman and Tanase 2018, 333)	259
Figure 5.16 A ResearchSpace knowledge graph represents data in a network of meaningful relations (https://researchspace.org , accessed 7 September 2024)	261
Figure 5.17 The three layers of a semantic knowledge graph (https://blog.metaphacts.com/importance-of-semantic-knowledge-graph , accessed 7 September 2024)	262
Figure 5.18 A user interface in metaphactory (https://blog.metaphacts.com/visual-ontology-modeling-for-domain-experts-and-business-users-with-metaphactory , accessed 7 September 2024).	263
Figure 5.19 The Golden Agents consortium (https://www.goldenagents.org/ga-output , accessed 7 October 2024).....	274
Figure 5.20 The Golden Agents schema – Project Proposal pp 14 https://www.goldenagents.org/about/	275
Figure 5.21 The Golden Agents – zooming in on one event (https://www.goldenagents.org/ga-output , accessed 7 October 2024).	276
Figure 5.22 The ROAR++ ontology (van Wissen, Zamborlini, and van den Heuvel 2022).....	277
Figure 5.23 The Golden Agents linked datasets universe (https://www.goldenagents.org/ga-output , accessed 7 October 2024).....	278
Figure 5.24 ECARTICO in the Golden Agents (https://www.goldenagents.org/ga-output , accessed 7 October 2024).....	279

Figure 5.25 The Bredius Excerpts in Golden Agents Notary Deeds (Register of death goods) (https://www.goldenagents.org/ga-output , accessed 7 October 2024)	280
Figure 5.26 The Notary Network in Golden Agents (https://www.goldenagents.org/ga-output , accessed 7 October 2024).....	281
Figure 5.27 Lenticular Lens – matching names from multiple datasets (Idrissou, Van Wissen, and Zamborlini 2022).....	283
Figure 5.28 Disambiguating SAA indexes with ECARTICO (Idrissou et al. 2018)	285
Figure 5.29 Modules with levels of detail and uncertainties (van Wissen, Zamborlini, and van den Heuvel 2022)	286
Figure 5.30 Building provenance in Golden Agents: Step 1 – locating and identifying data (https://www.goldenagents.org/ga-output , accessed 7 October 2024)	287
Figure 5.31 Building provenance in Golden Agents: Step 2 – linking up marriage data (https://www.goldenagents.org/ga-output , accessed 7 October 2024)	287
Figure 5.32 Building provenance in Golden Agents: Step 3 – linking up locational data (https://www.goldenagents.org/ga-output , accessed 7 October 2024)	288
Figure 5.33 Result of Golden Agents Query for endurant name = Rembrandt (https://ga-wp3-qb.sd.di.huc.knaw.nl , accessed 7 October 2024)	289
Figure 6.1 A Quaker Led Group topology	340
Figure 6.2 Visual Studio Code was used to build the database	344
Figure 6.3 DBeaver Desktop.....	345
Figure 6.4 The Project Seven GitHub repository	345

Figure 6.5 A Project Seven JNB	346
Figure 6.6 Gephi Visualisation Tool displaying the raw Project Seven data.....	347
Figure 6.7 Example of a container for Project Seven.	348
Figure 6.8 The HDDT ERD.....	349
Figure 6.9 RAI data Thomas Hodgkin1 part 1	352
Figure 6.10 RAI data Thomas Hodgkin1 part 2	353
Figure 6.11 RAI data Thomas Hodgkin1 part 3	354
Figure 6.12 Sample APS data collection in Excel	355
Figure 6.13 Quaker Committee on the Aborigines (four Excel workbook sheets).....	356
Figure 6.14 Quakers	357
Figure 6.15 Quaker family relationships.....	357
Figure 6.16 Data cleaning reconciliation	361
Figure 6.17 Image of the First Annual report of the (Aborigines Protection Society 1838)..	363
Figure 6.18 Image of London Yearly Meeting 1832 – the names of members of the Quaker Committee on the State of the Heathen (part of the QCA)	364
Figure 6.19 CEDA memberships by society	369
Figure 6.20 The person file	371
Figure 6.21 Correcting date errors.....	372
Figure 6.22 Correcting Quaker records.....	372
Figure 6.23 Data View sample	373

Figure 6.24 Data Views were audited by using the count function (1)	374
Figure 6.25 Data Views were audited by using the count function (2)	374
Figure 6.26 Gephi options set-up	376
Figure 6.27 The relationships in the CEDA database	377
Figure 6.28 Members of the CEDA	379
Figure 6.29 Locations pie chart.....	380
Figure 6.30 Networks span many nations	381
Figure 6.31 The CEDA network with the QCA marked in green and Hodgkin's network marked in red.....	385
Figure 6.32 Archivist's note on the QCA Friends House Archives	387
Figure 6.33 QCA joiners by year.....	388
Figure 6.34 QCA leavers by year	389
Figure 6.35 APS joiners in each year	390
Figure 6.36 APS leavers in each year	390
Figure 6.37 APS Quaker joiners in each year	391
Figure 6.38 APS Quaker leavers in each year	392
Figure 6.39 All Quakers and their family relationships – circled in blue are no or only one relationship	396
Figure 6.40 Filtering graphs in Gephi.....	397
Figure 6.41 Quaker relationships by type.....	399

Figure 6.42 Quaker presence in the CEDA.....	400
Figure 6.43 Distant relationships	401
Figure 6.44 Close relationships	402
Figure 6.45 Immediate relationships	404
Figure 6.46 Thomas Hodgkin's family relationships	405
Figure 6.47 John Hodgkin's family relationships	406
Figure 6.48 Edward Backhouse's family relationships	407
Figure 6.49 William Fowler's family relationships	408
Figure 6.51 The CEDA with 'ZOË' and 'WEL'	416
Figure 6.52 Quaker families in the HOD network.....	417
Figure 6.53 The emergence of smaller groups	418
Figure 6.54 Hodgkin apart from 'ZOË' and 'WEL')	419
Figure 6.55 The emergence of key individuals	420

Tables

Table 1.1 Metadata taxonomies, taken from the Dictionary of Archives – Terminology (https://dictionary.archivists.org/category/basic-archival-science.html , Accessed 25 July 2024)	15
Table 2.1 The CEDA database records (https://www.w3.org/RDF , Accessed 13/09/2023) ...	43
Table 2.2 Records and their digital representations	47
Table 3.1 The dominance of literature, languages and linguistics in DH (Sula and Hill 2019, 199)	82
Table 3.2 Source and tool critiques (Koolen, Van Gorp, and Van Ossenbruggen 2019, 374)	128
Table 3.3 Timeline of the history of digital databases ((Berg, Seymour, and Goel 2013, 30- 33))	135
Table 4.1 Implementing ICA Records in Contexts-Ontology (RiC-O) at the National Archives of France (ANF): first steps and prospects (table built from bullet points on a PowerPoint slide in (Clavaud 2021))	159
Table 4.2 Percentage of citations found by each database, relative to all citations (first row), and relative to citations found by the other databases (subsequent rows) (Martín-Martín et al. 2021, 882)	167
Table 4.3 FRBR founding objectives (International Federation of Library Associations and Institutions 2009, 8)	175
Table 4.4 Metadata types (Riley 2017, 7)	183

Table 4.5 Problems in using archival metadata vocabularies in the LOD environment Karen Coyle and Emmanuelle Bermes, W3C, Cluster Archives, September 2011 (https://www.w3.org/2005/Incubator/ldl/wikil/Cluster Archives , quoted in (Willer and Dunsire 2013, 267).....	185
Table 4.6 Extracts in tabular form (International Council on Archives 2004, 1.7)	204
Table 4.7 Entity 21 Person in the CIDOC CRM (Bekiari 2021, 73-74)	215
Table 5.1 Number of I-CeM downloads. Data supplied by the I-CeM team in response to a Freedom of information request by the author	250
Table 5.2 Number of I-CeM Special Licences. Data supplied by the I-CeM team in response to a Freedom of information request by the author	251
Table 5.3 Chapter 5 project review table	292
Table 6.1 Database specification	351
Table 6.2 Comparison of persons PEH in Case Study 3	412
Table 6.3 Comparison of persons WEL in Case Study 3.....	413
Table 6.4 Case Study 3 data sources.....	414
Table 6.5 Project comparison table	423

Chapter 1 The digital study of Past Human Lives

1.1 The Digital Humanities background

This thesis addresses a common concern in the rapidly developing field of Digital Humanities (DH) (Castells 2011) (Hilbert and López 2011) that the take-up of DH research practice is lower than might be expected given the relatively recent and enormous efforts put into digitisation in the field. Is this because DH researchers, and especially historians, are too tradition-based and so reluctant to adopt new practices? Is it because the new digital practices are too young, and therefore as yet unsettled, with new technologies both too complex and emerging too rapidly for researchers to feel confident in their use? Perhaps the promise of new technologies is oversold, and the possibilities of the new practices are over-promised?¹

This thesis takes an information science based approach to these questions about the study of past human lives and also in formulating recommended actions in Digital Humanities to address them.² (Bawden 2015) The central problem addressed is that digitisation so far has

¹ Discussed at length in Champion (2016).

² The thesis recognises the definition of Information Science developed by Harold Borko in 1968 when the American Documentation Institute voted to change its name to the American Society for Information Science. Borko argues that modern information science is born out of library science and his forward looking definition anticipates today's digital world - 'Information science is that discipline that investigates the properties and behaviour of information, the forces governing the flow of information, and the means of processing information for optimum accessibility and usability. It is concerned with that body of knowledge relating to the origination, collection, organization, storage, retrieval, interpretation, transmission, transformation, and utilization of information. This includes the investigation of information representations in both natural and artificial systems, the use of codes for efficient message transmission, and the study of information processing devices and techniques such as computers and their programming systems. It is an interdisciplinary science derived from and related to such fields as mathematics, logic, linguistics, psychology, computer technology, operations research, the graphic arts, communications, library science, management, and other similar fields. It has both a pure science component, which inquires into the subject without regard to its application, and an applied science component, which develops services and products.' (Borko 1968, 3)

not comprehensively revealed the evidences of all past human lives (too many lie locked in the bodies of documents held at archive, and little effort has been made to identify these embedded individuals or to disambiguate evidence related to persons, especially when linking records across archives. Furthermore, little work has been done to integrate archival with non-archival evidence, such as that found in genealogical records and their affordances.

The solution proposed here is to take advantage of,

- the opportunities presented by the rise of the relational Semantic Web,
- the major expansion of digital processing capacities and capabilities in the humanities over the last 50 years,
- the continued advances in GLAMS digitisation and metadata interoperability (including standards consolidation),
- the rise of genealogy as a major interest outside of the academy,
- and an emerging interest in all past human lives across an ever widening academic, professional, lay and amateur researcher community.

Because of these advances, what was impossible to achieve only a few years ago, is now possible.

The place where all of these advances meet is in Digital Humanities and it is there that significant further advances can now be made using information science to develop a methodological approach to the digitisation of the physical evidences of past human lives, and the careful management of their digital referencing and representation. This can be achieved through the adoption of an approach to the digital study of past human lives based on evidence based prosopography, and a related new National Authority Indexing system

for the digital management and ordering of prosopographical data on all past human lives.

This is a scientific and methodological approach to the systematisation and management of the evidences of past human lives and it aims to service the whole researcher community, in and beyond academia.

In academia, a lot of digitising activity has already taken place at archives and in collections over the last fifteen years or so, and these include the digitisation of finding aids. There has been a lot of thought given to how individual collections can be made digitally available to researchers through metadata. Comparatively less thought has gone into what researchers in the future might actually need, want or be inspired to do, when working with data alongside, or instead of, the physical things at archives that contain information about Past Human Lives (PHL). A rethink of theory in DH, resulting from the practical exercises undertaken for this thesis, calls for a move away from the current focus in DH on technologies and the digitising of legacy finding aids, towards building an infrastructure system to facilitate the finding, digitising and managing of Evidence Based Prosopographical (EBP) information and its representation as data. Jonathan Blaney, Jane Winters, Sarah Milligan and Martin Steer in *Doing digital history: a beginner's guide to working with text as data* ask the Question, 'If you cannot read everything, and keyword searching produces an overwhelming volume of results, how else can you think about analysing your sources?'³ They posit an answer calling for researchers to become ever more engaged with technology.

³ 'We expect that new ways of delineating, analysing and representing digital data will increasingly be used by historians. The growing popularity of network analysis, for example, is indicative not just of research interest in the connections between people, places and organisations but of data which is too large and complex to be viewed and presented in more traditional ways. If you cannot read everything, and keyword searching produces an overwhelming volume of results, how else can you think about analysing your sources? We already rely on search algorithms that we do not [and perhaps cannot] really understand, but we believe that greater reliance on artificial intelligence in research will require much deeper engagement with the technology. Historians will need to understand the tools that they are using, not just the sources that they are working with' (Blaney et al. 2021, 154).

This thesis argues instead that a better answer is to improve access to important information in records of Past Human Lives through an increase in permissive structure, organisation and affordance for EBP information when it is digitally represented as data. Improvements to the research environment and infrastructure provision, and in particular access to information on PHL and its digital representation will enable DH to take full advantage of the possibilities offered by the 'Semantic Web' and especially relational data systems (Berners-Lee Tim, Hendler James, and Ora. 2001); (Zandhuis 2005); (Meroño-Peñuela et al. 2015); (Allemang and Hendler 2011b). An EBP system that takes full advantage of the Semantic Web is proposed that will lead to better services for researchers and, aspirationally, even encourage a change in how society views, understands, preserves and uses information and representative data on PHL.

The universal digitisation of EBP present in the Records⁴ of PHL as Evidence Based Prosopographical Information (EBPI) is proposed as a new and necessary infrastructure affordance in digital research into PHL. EBP manifests as abundant information present in the Records of PHL and because EBP is naturally structured information it can be readily represented in digital form as data (EBPD). EBP is already partially present in the digital affordances of both archives and genealogical platforms. This thesis will show that by adopting a systematic approach to the finding and use of EBP, it is possible to bring together the EBPI pointed to in the digital platforms of academia and the EBPD already present in genealogical platforms, and to forge them into one virtual integrated digital affordance, whilst at the same time protecting the academic sensibilities of the one, and the commercial sensitivities of the other. This affordance would greatly improve the extent to which

⁴ See Section 1.2 for a definition of key terms.

digitisation achieves the universal and structured representation of EBPI on PHL as EBPd.

The adoption of the EBP system explained here would allow the relationships between EBP instances in the Records of PHL to be discovered and made durably accessible, efficiently managed and fully interoperable as data. What EBPI is, how its digital representation as EBPd should be structured, organised and managed, and how EBPI and its related EBPd can be brought into a durable and sustainable relationship are the subjects of this thesis. In devising an EBP systematology, simplicity in both concept and application was important, and simplicity has been a guiding principle for this thesis.⁵

The habitual piecemeal focus on discrete sets of data found in recent DH academic research projects to date neglects the need to think more widely and deeply about information on PHL at the highest level and how it should be digitally represented. Genealogical platforms are already significantly developed as digital affordances of EBP and they largely conform to the EBP system explained here. For this reason genealogical platforms are not studied here in detail although they are a major component in the EBP system. This thesis instead focuses on a detailed scrutiny of how academia has failed to fully learn from genealogy, and how current digital affordances of EBP in academia, while they are in place, require significant

⁵ 'The steep learning curves associated with software and data, especially those required to fully engage in reproducible research workflows, represent real obstacles to many archaeologists at all career stages. The complexity barriers to reuse may highlight tensions between the interests of data creators and data reusers. Data reusers will want relatively simple, consistent data at a scale and significance that make investing time in understanding and analysis worthwhile. Scale and consistency can best come when data conform to common standards. On the other hand, data creators typically want freedom to adapt and shape data recording practices to meet their immediate needs, ingrained habits, and particular research agendas. This can result in a great deal of variety and inconsistency in how data are created. Currently, digital repositories try to reduce barriers for data creators, by making deposit as easy as possible. However, the lack of consistency between datasets makes it harder for repositories to meet the needs of data reusers. Achieving greater interoperability (consistency) across diverse datasets comes at the cost of greater complexity. In the case of tDAR [tDAR is an online archive for archaeological and historical preservation information], that interoperability requires ontology mapping together with sophisticated software. In Open Context's case, interoperability comes about through schema mappings and annotations to common controlled vocabularies and ontologies via linked open data. Neither approach is simple or broadly understood by the larger archaeological research community' (Kansa and Kansa 2018, 4).

modification to allow the EBP system to take full advantage of the EBPI present in the Records of PHL held in academic archives. Commonly, academic digital affordances point to, but do not represent, EBP present in the Records of PHL. The disciplined and methodological approach to the management of both the physical instances of EBPI and their EBPD digital representations in archives⁶ set out in this thesis can facilitate the linking of EBPD in archival Records to EBPD in genealogical records. This will lead to new and better ways of research practice, and a widening of research interests, in the study of PHL.

In terms of structure, this thesis takes its inspiration from Harold Borko, *Information Science: What Is It?* (Borko 1968):

Information science is that discipline that investigates the properties and behaviour of information, the forces governing the flow of information, and the means of processing information for optimum accessibility and usability. It is concerned with that body of knowledge relating to the origination, collection, organization, storage, retrieval, interpretation, transmission, transformation, and utilization of information. This includes the investigation of information representations in both natural and artificial systems, the use of codes for efficient message transmission, and the study of information processing devices and techniques such as computers and their programming systems. It is an interdisciplinary science derived from and related to such fields as mathematics, logic, linguistics, psychology, computer technology, operations research, the graphic arts, communications, library science, management, and other similar fields. It has both a pure science component, which enquires into the subject without regard to its application, and an applied science component, which develops services and products, (Borko 1968, 3).

Borko is concerned with the life cycle of information and applied to this thesis this means Evidence Based Prosopographical Information from its instantiation in Records of Past Human Lives through its representation in forms of digitisation, as metadata, and in its use

⁶ '[I]t is important to understand that all descriptions in archives are information representation, which in turn are a form of representation. Rosenberg argues that "the essential and characteristic human activity is representation—that is, the production and manipulation of representations" (Rosenberg 1981, p. 1). Representations have certain philosophical constraints that emanate throughout the represented objects. One is that a representation is, at its fundamental level, "something that stands, or is believed to stand, for something else" (Yeo 2018, p. 129)' (Pacheco, Da Silva, and De Freitas 2023, 635).

in research activities. It has a pure science aspect, taken up here by sections that define the nature, qualities and concerns in the use of Evidence Based Prosopography throughout its various forms, and how these concerns have been neglected because of the dominant focus on technologies and tools instead of on information and data; and then an applied science aspect, taken up here in a practical research project (Project Seven – online and at Appendix 7)⁷ where EBP was used taken from a wide range of information manifestations. The thesis sets out issues in inherently messy data complexity, including provenance, fixity and affixedness. Structural issues such as the need for researchers to work seamlessly with both open-source and commercial datasets, and issues in the development of digitisation in academic archives, especially in metadata provision and metadata standards, are considered. The thesis urges the adoption of an independent National Authority Index and the systematisation of the use of Evidenced Based Prosopography in all researches into Past Human Lives because these bridge between the academic open-source affordances in academia and the commercial property rights of genealogy and other commercial owners of EBPI.

This chapter first introduces the key terms used in this thesis (Section 1.2). Section 1.3 introduces the practical research project I undertook to support the theoretical focus of this thesis, Project Seven (P7). The critical assessment of infrastructure provision in DH for the study of PHL considered here could not have taken place without a thorough understanding of research practice gained in the design and building of a Human Data Digital Toolkit (HDDT) and using it to address a valuable research project into the EBP relationships between 3000 activists from 1830 to 1870 (see Chapter 6). The concerns that arose from

⁷ <https://kelvinbeerjones.github.io/project-seven-book/intro.html> (Accessed 19/04/2025)

undertaking the P7 project are introduced in Section 1.4. Section 1.5 introduces the concept of EBP as it is applied in this study. Section 1.6 then outlines the National Authority Indexing–Unique Identifiers (NAI-UID) system, which is a key component of the proposed EBP system. The objectives of this thesis are then set out in Section 1.7 and finally the thesis chapters are outlined in Section 1.8.

1.2 Definitions of key terms

The following special terms, definitions and initialisms occur throughout this thesis. They are arranged in a logical flow from the objects the thesis considers to the approaches to their study.

PHL. Past Human Lives – persons no longer living who are not covered by applicable data protection laws and other prohibitions, who can be identified with a person name. All humans have names, therefore nearly all EBPI information found in Records of PHL can be identified by person names. However, person names are not fixed in form and are not unique, therefore they can be notoriously difficult to work with. The proposed EBPD system must therefore be flexible and permissive, allowing allocations to the NAI-UID indexing system to be made based on the judgement of the archivist/researcher to accommodate variability and uncertainty in the identification of person names and in their relationships.

Record (of PHL). Any physical thing or item held in an archive or other academic collection that contains or bears information about PHL.⁸ Records are also called here (see for instance

⁸ 'Researchers might treat audio recordings, images, physical and biological specimens, or textual documents as data. For instance, primary source documents are central to historians' work. Using machine-learning

Section 4.4), 'items', 'things' and 'documents'. The term 'Records' as used in this thesis, in particular embraces the term 'documents' set out by Paul Otlet in *International organisation and dissemination of knowledge: selected essays of Paul Otlet*:

Documentation is understood to mean bringing into use all of the written or graphic sources of our knowledge as embodied in documents of every kind, though chiefly printed texts. These documents consist of whatever represents or expresses an object, a fact. or an impression by means of any signs whatever (writing, picture, diagrams, symbols)... In a general way, one can say that documents of all kinds, the production of which began centuries ago and continues unceasingly in all countries, are registering or have registered. day by day, all that has been discovered. thought, imagined, planned. Thus, they constitute the means by which all of this has been transmitted from generation to generation and from place to place. As a whole, then, documents form the graphic memory of humanity, the physical body of knowledge, (Otlet and Rayward 1990, 115).

For Otlet a document was a physical thing, an entity in itself. Importantly for this thesis, and the reason why Record was chosen over Document is that a record is any physical thing that an observer accepts as a vehicle which carries information about Past Human Lives.⁹

Prosopographical information (information). Prosopography, the collection and interpretation of information on PHL, examines the social context of wide groups of connected individuals. Prosopographical information in this thesis is the incidences of

models, computational social scientists might analyze government reports, white papers, news articles, or other information. Archeologists treat fossils as data to analyze evolutionary patterns. However, we argue that what ultimately matters is not the format or package within which data are contained (e.g., textual documents, images, audio recordings, quantitative files). Rather, what matters "data" is that are a type of information that serves as an object of analysis, which may possess different research affordances' (Million et al. 2024, 651).

⁹ 'The notion of a document is central to library and information science. Library and information work, including knowledge organization, centrally involves the creation, processing, and organization of documents. Library and information science scholars have thus developed sophisticated understandings of what it means to call something a "document." Michael Buckland (1997; 2014) outlines how particular entities can be "made as," "made into," and "considered as" documents. These three views, which are progressively more inclusive, reflect how particular things may be; 1) deliberately designed to serve documentary purposes ("made as documents"); 2) human artifacts may be used as documentary resources even if that was not their original purpose ("made into documents"); or, 3) naturally occurring objects such as rocks or animals may be used for documentary purposes ("considered as documents"). In these senses, almost any object could be used as a document depending on their evidentiary value in particular circumstances. Being a "document" is, therefore, a role that particular things play, rather than an inherent property of those things. (Mayernik 2021, 701-702).

person names, person attributes and the descriptions of relationships between persons found in Records. The definition is explored in depth in Section 2.2.

Evidence (of PHL). Evidence is the occurrence of any incidence contained in a Record which is judged to be evidence of PHL. King's College London has a related but different approach to prosopographical information taken here that ascribes to the incidence itself the quality of 'factoid'.¹⁰ A comparison of Evidence as it is used here and 'factoid' where 'factoid' describes a quality of the information itself – it either is, or it is not, a 'factoid' – whereas in this study the term Evidence is used because it locates the definition of the term in the observer and not in the incidence. Defining an incidence of information as Evidence requires a judgement on the part of the researcher that the incidence can be taken as Evidence.¹¹ In this thesis, Evidence is not a quality of the item itself, it is a judgement of the observer.

Evidence Based Prosopographical Information (EBPI). Evidence Based Prosopographical Information must be (1) prosopographical in nature, (2) capable of being represented in

¹⁰ 'A factoid is not a statement of fact about a person; a collection of factoids does not record a "scholarly overview" of a person that a scholar has derived from the sources s/he has read. Instead, each one records an assertion by a source at a particular spot about a person. Factoids may contradict each other—if one source says a person was an Armenian, and another that s/he was Bulgarian, then both factoids will be present in the database. The ironic flavour of the name 'Factoid' is not accidental. It reflects the historian's worry when a tiny extract is taken out of the context of a larger text and the historical period in which it was written and presented as a "fact" (J. Bradley and Short 2005, 8). Also see <https://www.kcl.ac.uk/factoid-prosopography> (Accessed 20 March 2025).

¹¹ 'Multiple lines of recent work have developed conceptual definitions of 'data' that are built on the notion of evidence. Borgman (2015), synthesizing a series of ethnographic studies of science, social science, and humanities research, conceptualizes data as 'entities used as evidence of phenomena for the purposes of research or scholarship'. Similarly, Leonelli (2015) define data 'as a relational category applied to research outputs that are taken, at specific moments of inquiry, to provide evidence for knowledge claims of interest to the researchers involved'; see also Leonelli, 201 6a). On this view, researchers do not generate or collect 'data'. They instead generate or collect entities, which might include physical objects, measurements, or other inscriptions, that can be used as data in relation to specific research goals. Being data, on this interpretation, is a rhetorical and sociological role that entities are made to play in particular situations (Rosenberg, 20 13), not an inherent property of those entities that can be divorced from circumstances of their use' (Mayernik 2019, 733-734).

digital form as EBPB, (3) attached to a person name, and (4) capable of matching to person names in the NAI-UID, by the exercise of the researcher's judgement.

Evidence Based Prosopographical Data (EBPD). Also called **representative data**. Instances of EBPI rendered in a digital format are EBPB. The relationship between incidences of EBPB and the unique incidence of the EBPI it points to should include the quality of fixity (changelessness) and affixity (securely attachment).¹² All EBPB is one part of a bound dual relationship, with EBPI as its attached referent.¹³ EBPB, because it includes a person name in its manifestation, can be used to form indexes organised by person name and like EBPI, EBPB can have attributes attached to it, allowing for extensibility in the NAI-UID system. EBPB can be disseminated over the Web by the process of copying from approved (by some reputable authority) instances of EBPB allowing for provenance tracking in the NAI-UID system. This minimises possible corruption of data through dissemination.

Unique Identifier (UID). This has the meaning set out in *Understanding metadata* by Jenn Riley:

¹² 'Historians have long had access to source commentaries describing the characteristics of certain types of historical sources, like parish registers of baptisms, weddings and funerals, legal deeds and estate inventories, to name just a few. As an extension of this, it must be possible to compile generic data models for sources – in the same way as tei [Text Encoding Initiative] has designed generic dtDs [Document Type Definition Systems] for literary texts. The term "generic" refers to the data structure shared by different variants of the same main source type, and to a basic functionality which characterises the source type in question. If such generic data models were available to us, data models for specific research exercises would be quicker to produce, and the resulting data collections easier to share' (Breure, Doorn, and Boonstra 2006, 107).

¹³ 'One of the structuring principles of this dissertation is that digital artifacts, that is, digital media objects made by archaeologists, need to be understood as active members of a network of interpretive meaning. This active membership in the construction of meaning must be emphasized as digital objects are often imagined to be ephemeral or insubstantial, existing somewhere "in the cloud." To understand the relationship between an artifact or an archaeological site and its digital simulacrum, we need to understand that these objects are substantial in their own right. While they have different qualities and affordances, the digital object can be understood and studied not just as a reflection, a snapshot, or an echo of its real counterpart, but as an artifact with its own attending network of meaning. To this end, this chapter describes the understanding of materiality generally and within archaeology, adding the background of Visual Studies to aid in interpretation of digital materials, considers the misperception of digital media's immateriality, and finally, grounds digital materiality within archaeology' (Morgan 2012, 7).

It is Linked Data best practice for URIs to be dereferenceable. This means that URIs should be actionable via HTTP so that both humans and machines can see useful information such as labels, definitions, and relationships to other resources when visiting the HTTP URI for an RDF-encoded concept. The process of content negotiation is used to provide human users with Web pages and software applications with raw data when they visit the same URI. This allows RDF-aware software applications to make use of classes and properties with which they are unfamiliar, and makes RDF-encoded Linked Data a powerful tool for connecting information from multiple Sources. (Riley 2017, 11)

National Authority Indexing System (NAI-UID). A national digital indexing system of person names based on national open-source birth, marriage and death (BMD) records. For example, the NAI-UID for the UK could be based on one now being indexed with UIDs by the General Records Office (GRO).

Archival Authority Index (AAI-UID). An archival digital indexing system of person names to record person names present in archival records where each unique person name is allocated a digital UID by the institution holding the record. This is then cross-referenced to the National Authority Index on a best fit/probability basis (UID matching). The AAI-UID could now be based on the UIDs allocated to catalogue, index and finding aids data in current archival digital affordances.

Researcher Authority Index (RAI-UID). A research project digital indexing system of person names to record unique person names present in each research project database. Each unique person name present is allocated a digital UID by the researcher responsible for the digital EBPD related project. This is then cross-referenced to the National Authority Index on a best fit/probability basis (UID matching). If the EBPI or EBPD was taken from an archive, then that archive's AAI-UID is also recorded to facilitate the tracking of information to its original record.

Birth, marriage and death (BMD) records. National or customary systems for the official recording of births, marriages and deaths. These are frequently ecclesiastical records. BMD records, together with census data, are the bedrock of EBPD provision on genealogical platforms. In the UK, Free UK Genealogy hold BMD indexes.¹⁴

Galleries, libraries, archives, museums and special collections (GLAMS) metadata.

Metadata in GLAMS consists of the catalogues, indexes and finding aids developed and used primarily by GLAMS staff to describe, locate and organise the Records in their care. This metadata is often linked between GLAMS organisations. The metadata has different characteristics in each sector, resulting in different metadata standards, but efforts at integrating standards are now taking place (see Chapter 4).¹⁵

Metadata. This has the meanings set out in Table 1.1.

Data	<p>Facts, ideas, or discrete pieces of information, especially when in the form originally collected and unanalysed.</p> <p>Traditionally a plural noun, data - rather than datum - is now commonly used with a singular verb. Data often is used to refer to information in its most atomized form, as numbers or facts that have not been synthesized or interpreted, such as the initial readings from a gauge or obtained from a survey. In this sense, data is used as the basis of information, the latter distinguished by recognized patterns or</p>
------	--

¹⁴ <https://www.freeukgenealogy.org.uk/> (Accessed 18/04/2035)

¹⁵ 'Libraries take a "bibliographic" approach to metadata, which is rooted in their traditional strength in describing books. Bibliographic metadata focuses on detailed descriptions of individual items that allow users to locate these items. Archives use "finding aids," descriptive inventories of collections, along with historical information necessary for understanding the material. For archives, metadata helps users locate groups of related items that arise from the regular work of an individual (which are called papers) or organization (which are called records), and that contain material best understood in the context of that grouping' (Riley 2017, 5).

	<p>meaning in the data. The phrase 'raw data' may be used to distinguish the original data from subsequently 'refined data.' Data is independent of any medium in which it is captured. Data is intangible until it has been recorded in some medium. Even when captured in a document or other form, the content is distinct from the carrier.</p>
Metadata	<p>Information about data that promotes discovery, structures data objects, and supports the administration and preservation of records. Metadata may be embedded or external. It may be applied at a variety of levels of granularity and during different periods in the life cycle of data. It is typically demarcated and standardized, and it often provides context.</p>
Administrative metadata	<p>Data that is necessary to manage and use information resources and that is typically external to informational content of resources. Administrative metadata often captures the context necessary to understand information resources, such as creation or acquisition of the data, rights management, and disposition.</p>
Descriptive metadata	<p>Information that refers to the intellectual content of material and aids discovery of such materials.</p> <p>Notes: Descriptive metadata allows users to locate, distinguish, and select materials on the basis of the material's subjects or 'aboutness.' It is distinguished from information about the form of the material, or its administration.</p>
Preservation metadata	<p>Information about an object used to protect the object from harm, injury, deterioration, or destruction.</p>
Structural metadata	<p>Information about the relationship between the parts that make up a compound object.</p>

Table 1.1 Metadata taxonomies, taken from the Dictionary of Archives – Terminology
(<https://dictionary.archivists.org/category/basic-archival-science.html>, Accessed 25 July 2024)

Human Data Digital Toolkit (HDDT). The assemblage of digital technologies used by a researcher to collect, organise, analyse and visualise EBPD.

1.3 Project Seven (P7) introduction

The theoretical aspects of this thesis are underpinned by practice. I identified a group of people who in the nineteenth century worked together over many years (1830–1870), initially to champion the rights of aborigines throughout the British colonies, and who later enabled the disciplinisation of anthropology in Britain. Around 3000 people altogether, there were around 600 Quakers among them. Prominent among the Quakers and the wider group was Thomas Hodgkin MD (1790–1866). His central organising role emerges from the P7 study.

This group of 3000 people were all members of:

- The Quaker Committee on the Aborigines, QCA (1831–1846)
- The Aborigines Protection Society, APS (1837–1848)
- The Ethnological Society of London, ESL (1843–1848)
- The Anthropological Society of London, ASL (1861–1869)
- The Anthropological Institute, AI (1871), a merger of the ESL and the ASL

Collectively, these societies are called the Centres for the Emergence of the Discipline of Anthropology in Britain (CEDA). A DH project was devised to analyse and visualise the

relationships between these people in the form of an archival database using EBPI contained in a variety of primary sources and metadata. The Project Seven Report appears as Appendix 7. The P7 project was tasked with answering three questions:

Question 1

Can the model reveal the networks between the members in the five organisations that comprise the CEDA? This question is important because it resolves a wider and current uncertainty over the origins of the discipline of anthropology in Britain and the extent of Quaker involvement.

Question 2

Can the model examine Quaker-to-Quaker relationships and how these relationships supported the Quaker members of the CEDA during the forty years of its life?

Question 3

Can the model reveal the key networking role played by the Quaker Thomas Hodgkin MD (1798–1866) from the beginnings of the CEDA in 1830 up to his death in 1866?¹⁶

The work of researching manuscript and microfilm sources and combining EBP extracted from them digitally with other information in the form of metadata produced a P7 dataset. I then designed and built a suite of technologies to form a digital research tool to clean, manage, organise, analyse and visualise the data. This suite of technologies is called the Human Data Digital Toolkit (HDDT). Claire Warwick, in 'Building theories or theories of

¹⁶ There are several secondary sources that cover the life of Thomas Hodgkin MD. Two that this project will reference frequently are *Perfecting the world: the life and times of Dr. Thomas Hodgkin 1798–1866* (Kass 1988) and *Protecting the empire's humanity: Thomas Hodgkin and British colonial activism 1830–1870* (Laidlaw 2021).

building? A tension at the heart of digital humanities’, explains the importance of practice in support of theory in DH:

We cannot fully understand digital resources until we have made them ourselves: that it is only by writing programmes, building databases, digitising material or marking up text that we fully appreciate the complexity of what might appear relatively simple in digital terms, at least to the non-maker [Turkle 2009]. This philosophy lies behind much of the teaching that we do in DH programs. We do not necessarily expect our students to be full time digitizers, encoders, or programmers, but we teach them to do so, so that they should understand the digital objects that they will go on to work with. (Claire Warwick 2015, 542)

The Project Seven report provides a description of the project and demonstrates the experiential learning and skills necessarily acquired to then undertake the theoretical half of the work which constitutes the majority of this thesis.

1.4 Researcher concerns identified in P7

When, after research activity is complete, a researcher places their own database (or just the dataset) into an archival repository for the benefit of other researchers, problems then arise. If there are poor or unhelpful standards in the study of PHL, or a lack of common methodologies, ontologies and other infrastructures to support and guide the work of researchers, then it will be very difficult for the individual researcher’s work programmes and work practices to be of practical use, even if researchers try to diligently follow FAIR standards.¹⁷ This concern arises because there are many technologies that researchers can use as research tools, many ways of building databases and many ways to use digital

¹⁷ ‘GO FAIR is a bottom-up, stakeholder-driven, and self-governed initiative that aims to implement the FAIR data principles, making data Findable, Accessible, Interoperable and Reusable (FAIR). It offers an open and inclusive ecosystem for individuals, institutions and organisations working together through Implementation Networks (INs).’ <https://www.go-fair.org/fair-principles> (Accessed 20 March 2025).

models, all of which make data conformity complicated and often lead to concerns over possible compromising of data quality. Many DH database-related research projects may on the surface look quite interoperable, but they are likely to be relatively individualistic on closer examination. This is especially so given the universal and fundamental qualities of complex, disorganised and often messy data that DH researchers routinely have to engage and grapple with. Consequently, unless the researcher's archived dataset is large (and so would be too time consuming to reinvent), it is likely that later researchers will prefer to revisit the original physical Records and start enquiries afresh. Once a research dataset is deposited in a repository, the likelihood of its reuse also diminishes over time, especially if the original research team are unavailable later to provide advice to the new team.¹⁸ How the original project objectives shaped the data and the many decisions taken that influenced the data outcomes are difficult to capture on data plans, on data sheets or in metadata. This concern can become acute if the dataset has over time become a lone surrogate of EBPI because the original EBPI source is now lost (see Section 7.5.5). There is also the lingering hope that new or more data might be found if the enquiry started afresh, which again weakens the resolve to reuse existing datasets.

This thesis argues that these important issues can be addressed by data scientists, genealogists and technologists working together, who would specialise in understanding and addressing the problems that arise at the interface between EBPI and EBPD, and how

¹⁸ 'While data sharing is still rare, it is gaining traction as a key issue in scientific communications (Costello 2009; Nature Editors 2009). Scholars have discussed a multitude of semantic (Kintigh 2006), technological (Snow et al. 2006), data preservation and longevity (Carraway 2011; Richards 2004), intellectual property (Kansa et al. 2005), and professional incentive concerns (Costello 2009; Kansa 2010) regarding data sharing. While most see data sharing as an important goal, much attention focuses on problems relating to supplying researchers with data and less on how researchers can best consume and reuse data. Despite wide acknowledgement that approaches to data collection, recording, analysis, presentation, and interpretation vary among researchers, few studies have explored challenges researchers may face in the analysis of datasets produced by others' (Atici et al. 2013).

these problems change over time. This thesis shows that only by bringing together the digital affordances for research into PHL currently found in both archives and genealogy through the EBP system, and then combining them into one integrated digital affordance, can a universal infrastructure and data provision be made to enable the comprehensive study of PHL.

1.5 Evidence Based Prosopography

Once the three P7 Case Study exercises were completed, it was possible to reflect on the experience of the design, build, use and value of the HDDT model itself, its associated technologies and methodologies, and the 'Independent Researcher' approach to research practice. This enabled a consideration of wider aspects of DH to better contextualise the Case Studies themselves, and for the theoretical part of the project to address the data and infrastructural concerns which arose. The following key learning points were identified:

- Much important EBP information on PHL (in the nineteenth century and earlier) lies undiscoverable in archives and collections because it has not been digitised, and where digitisation has occurred data is predominantly locally hosted and therefore not directly accessible via the Web.
- Much undigitised EBPI at archives might never become digitised, because digitisation of the Records of EBPI (the Records of ordinary human lives) is often considered of little relative value, largely because current funding attention is understandably focused on prestigious national digitisation projects.

- Metadata engines such as Dublin Core¹⁹ take document author as the key field and therefore do not accommodate the recording at catalogue levels of the EBPI person names that may be embedded in Records. Metadata in turn provides catalogue data to local and global digital search engines, though full text searching can usually only take place locally, working with one or a small selection of documents at a time.²⁰ For instance, Thomas Hodgkin MD may be mentioned in the body of many books, publications and manuscripts, but these mentions do not systematically appear in the metadata for those books.
- In the future there may be a failure to digitise EBPI. The rapid expansion of the digital world, for example with the advent of ‘machine learning’ and the shift to the ‘Semantic Web’ now well under way, risks the focus of future research efforts moving habitually away from EBPI and towards already digitised EBPD, leaving behind that which is not. Consequently, research into EBPI might soon become a niche, marginalised and specialised academic field. As a result, a hard fault line may emerge between the digital and pre-digital worlds, when that could be overcome if the digitisation of EBPI as EBPD became a core objective of DH and the relationships and linkages between EBPI and EBPD were systematised.
- The rapid deployment of digitisation to capture and organise EBPI is impaired because efficient use of digitisation for research purposes sometimes lacks the

¹⁹ ‘The Dublin Core Metadata Element Set (DCMES) grew out of a 1995 meeting in Dublin, Ohio, that was focused on metadata for networked electronic information. Attendees were tasked with identifying a core set of features common to most types of digital information. In this first meeting, 13 core elements were defined, which soon grew to the 15 elements known as DCMES today. These are: contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title, and type. This set, also known as “simple Dublin Core,” or Dublin Core (DC), is standardized as ISO 15836 and ANSI/NISO Z39.85, both called The Dublin Core metadata element set’ (Riley 2017, 23).

²⁰ The text of individual documents is not indexed by (for instance) Google ‘spiders’ and so content data does not appear in front-line searches.

unique identification and indexation needed to appropriately and structurally organise EBPI when it becomes EBP. ‘Structured’ and ‘organised’ here mean at least the indexing of birth, marriage and death certificates²¹ to form an NAI system to be used by all archives and researchers in their research projects.

- There needs to be an understanding that the physical and institutional separation of genealogists and archivists (and thus genealogical data from archive data) should end, so that all EBPI can be comprehensively studied and made available through its EBP representations. There is therefore a need for the marriage of genealogical and archival Records in order to capture and systematise linkable information as representative EBP. when this does not happen at the institutional level, it nonetheless does happen at the research hub/Academic or Other Researcher level, where each researcher or research team is currently obliged to construct their own local or project specific NAI equivalent. The current unstructured practice of localised authority indexing makes the future interoperability of multiple datasets uncertain.

EBP is prosopographical in nature and therefore it is science based; as such, it is concerned with information that is at least potentially verifiable contained in Records. For example, that Thomas Hodgkin attends Quaker meetings is information, but what he thought about his faith is not. Prosopography and its disciplines that form the core data items that make up the EBP system are well established in academic theory and practice, and methodologies for its use that will support the EBP system in research are already in place:

²¹ These are structured and verifiable, and all PHL data in archives can be referenced to one of these state-controlled documents. This is the only route to a comprehensive, trusted and universal authority index.

Prosopography is concerned with certain groups of people sharing certain common characteristics. The group is analysed through the study of its constituent parts, the different people who make up the set. So, at the heart of all prosopography lies the issue of identity, that is, the individualization of the separate persons in a mass of data relating to a group or groups – something it is much easier to state as obvious than to achieve in practice. The basis of a prosopographical dataset is an initial register of references to people occurring within the Sources being exploited. Such references are called name records, whether or not they include a personal name. (Keats-Rohan 2007, 151)

1.6 EBP and the National Authority Index system

The recognition that EBPI is ‘messy’ leads to the need to create an NAI system based on BMD records, because these are relatively universal and structured (if imperfectly) and they can provide a comprehensive framework with which to structure messy data. If a reputable, verifiable, universal authority index system is developed and put in place, then future researchers will be encouraged to add further EBP to the system and, through crowdsourcing initiatives and practices, extend and enrich the system by the inclusion of attributes through the extensibility features of such a system. Only then can EBP data science emerge as an important and central discipline in both DH and the information sciences.²²

This thesis proposes an NAI system based on Unique Identifiers (NAI-UID) and the NAI component of the EBP system is discussed at length in Chapter 2. Using the NAI system, research projects will be capable of full integration and so build up a common, shared, rich,

²² It is beyond the scope of this thesis (and the skills of the writer) to address here the concern about how to resolve the problems of handling ‘messy data’ in humanities sources and at the same time promote the preserving and revealing of all past PHL data, but it is right that these questions be asked. Favourable exploratory discussions have taken place with The National Archives and the General Records Office that indicate that the approach suggested here is of merit.

inclusive, virtually distributed, nationally based corpus of EBPB. This will facilitate the virtual integration of individual projects that over time could become a single national standardised EBPI to EBPB service. It is doubtful that researchers will feel confident combining distributed EBP datasets without the aid of a digital NAI system to provide a suitable research infrastructure.

Chapter 2 sets out the principal recommendation of this thesis: that all EBPI should be made fully discoverable and represented as EBPB, ensuring its digital usability in affordances that fulfil FAIR principles. To do so its associated EBPB should be organised through a system of authority indexes (see Section 2.6). At the highest level there should be one NAI-UID for each nation.²³ Each archive or collection containing EBPB, because it has a unique physical location, should have its own Archival Authority Index for the persons present in the sources at each institution (AAI-UID). The AAI-UID index can be linked on a best fit basis to the NAI-UID. There should additionally be a Researcher Authority Index (RAI-UID) created by each researcher or research team for each research project. The RAI-UID index should be cross-referenced to both the AAI-UID representing the location of the EBPI supporting the EBPB and the NAI-UID. The NAI allocation must be made permissively, on a person name best fit basis, to reflect the individual researcher's understanding and judgement of the identity of the person subjected to the matching process. In this system the NAI-UID constitutes the common link between all AAI-UIDs and RAI-UIDs.

For the NAI-UID system to work, genealogy must be invited into the DH 'Big Tent' (see Section 2.7). Deciding how to make links between AAI-UID and RAI-UID and the NAI-UID given the prevalence of incomplete and messy data will often be a matter of judgement. It is

²³ Official records of BMD, while occasionally collected by other institutions, are routinely and universally compiled and preserved at the level of the nation-state.

quite possible that a researcher and an archivist might wish to make different links between their AAI-UID and RAI allocations and the NAI-UID. Therefore, it is appropriate for the permissive record and evidence matching practices common among genealogical platforms to be used within both archival and researcher infrastructures. This will help minimise misallocations, and in turn allow the NAI-UID system to be flexible and permissive by allowing choices that reflect uncertainty and variability in name matching approaches and practices adopted between users.

In summary, the key components of the NAI-UID system are (1) the creation of a NAI-UID based on BMD information, (2) attaching UIs to every incidence of EBPI present in Records of PHL at archives and cross-referencing them to the NAI-UID, (3) researchers adopting the NAI-UID system and applying it to individual research projects by creating their own project-based RAI-UID, and then linking each Record to both the AAI-UID and the NAI-UID. The NAI-UID systems could encourage the development of an NAI-UID universal search engine (perhaps based on Google Scholar and its citation system) and this could then be expanded to accept the attachment of attributes – the NAI-UID system should be extensible. The EBP system should conform to and enable the Semantic Web structures of ‘Linked Data’ to enable relational connectivity between digital assets, UIs to uniquely identify those assets and a Resource Description Framework (RDF) to describe digital assets in the form of triples – for example, person-has-name UID, name-has-NAI Index UID, name-has-Archival Index UID, name-has-Research Index UID (see Figure 1.1).²⁴

²⁴ ‘The Semantic Web, it is argued, is the Web of meaningful data that can be processed by computers and employs “Linked Data” as the mechanism for publishing structured data on the World Wide Web where that data can be linked and integrated. It uses the same HTTP protocol [Hypertext Transfer Protocol] and a similar way of identifying data [Uniform Resource Identifiers, URI, or “web resources”], as that employed by web pages [W3C Technical Architecture Group, 2001]. However, in contrast to an HTML [HyperText Markup

National Authority Index	UUIDs	Location	
Thomas Hodgkin + DoB + M + F + Loc.	1234567890	GRO-UK	←

Archive Authority Index	Identifiers	Location	
Dr Hodgkin	2345678901	Physical at TNA	
Thomas Hodgkin NAI	1234567890	GRO - UK	←
Embedded in Document	Dublin Core UUID	Physical at TNA	

Research Database Authority Index	Identifiers	Location	
Quakers and activism	4567890123	Db a xxx repository	
Thomas Hodgkin MD - RAI	3456789012	Db key field	
Thomas Hodgkin NAI	1234567890	GRO - UK	←
Dr Hodgkin (Archive)	2345678901	TNA	
Embedded in Document	Dublin Core UUID	TNA	

Figure 1.1 Example of a National, Archival and Research Authority Index system based on BMD certificates

The NAI-UID concept (Section 2.6) is an essential part of the EBP system. Through it digital representations (EBPD) of instances of EBPI discoverable in archival Records can be fixedly, durably and affixedly connected to each other.

1.7 Thesis objectives

Five questions grounded in both theory and practice emerge from this study about EBP and the NAI system, and their roles in the future of digitisation in DH:

1. What is Evidence Based Prosopography and the National Authority Index system?

Are they the way forward for digitisation in the humanities?

2. Is infrastructure provision in the Digital Humanities sufficient to take up the national enterprise of the digitisation of Evidence Based Prosopography?

Language] Web page, the Web of Data uses a simple meta-model called RDF [Resource Description Framework] consisting of only three elements: a subject, a predicate, and an object, commonly known as a “triple” (Oldman, Doerr, and Gradmann 2015, 252).

3. Are digitised finding aids a good bridge to the records and the information contained in them that researchers are interested in?
4. How successfully have recent large-scale research projects into Past Human Lives used Evidence Based Prosopography?
5. How successfully has the small-scale Project Seven research project into Past Human Lives used Evidence Based Prosopography?

The thesis concludes that DH efforts in academia in the recent past have focused on digitising archival collections and Records as metadata, and on how digitisation can make Records more accessible to researchers, but EBP and how EBPI can be systematically represented as EBPD are relatively poorly serviced. What is needed now is serious reflection on meeting the digital needs of DH researchers – on the digital representations of EBPI.

Whether the EBP system or an equivalent should be adopted by DH is one of the most important issues facing DH today, because it systematises the relationships at the interface between the physical world of the Records of PHL and the digital world of data, thus linking the past to the present. DH in its first stages of development has often focused on adapting borrowed tools and digitising prestige collections and archival finding aids, when it should have also focused on using information science methods and methodologies to open up digital access to the EBPI contained in records.²⁵

²⁵ 'Others have described the role of the digital humanities or digital history as "ensur[ing] that digitization does not create black boxes, but meets scientific criteria of transparency, traceability and reproducibility." In a way, this results in digital historians modeling the research practices of their own discipline—history and history writing—in a way that resembles a new way of thinking but that is, in fact, what it always was, a mirror of the social condition in which history writing takes place. The models of knowledge in digital history refer back to the theories and traditions of historical research while digital historians also develop models for integrating the formalized thinking that is present in digital society. In doing this, their reflections on modeling and algorithms do not give way to new "black boxes" but rather help to open the "black box" of interpretation in which historians' deliberations, hypotheses, and conclusions (which are seldom explicated in their scientific work) take place' (Schwandt 2022, 84).

DH is still in its development phase and is subject to considerable uncertainty and change as the discipline grows and technology expands. At the same time, researchers from a variety of fields continue to (and must) rely on EBP in their research. This unsettled status of DH may persist for at least another generation. During this time important digital (and even physical) research data risks being lost. Furthermore, during this time the ability to work with physical historical information may fade, especially because 'Born Digital' research techniques will continue to expand exponentially. Therefore, over time, the transition to the digital world may mean that research using only physical information becomes a specialist and therefore niche practice.

Practical modelling shows that the focus of digitisation in DH has so far often been on data about research authors, academic publications and the borrowing of technologies from the sciences, at the expense of finding and understanding information and representing it as data. The Case Studies here urge that DH must now focus on digital affordances for researchers who need to use all primary information on past lives, wherever that information is found in archival Records. The EBP system seeks to maximise the digital affordance of the Evidence of PHL for research. In taking up this enterprise, DH will extend its reach beyond the arts into past human data science, bridging the gap between the humanities and the sciences in the digital world.

The concepts of EBP and its NAI system ensure that the Records of all PHL (and in spite of considerable effort in recovering lost records, some information will be irretrievably lost) are made digitally available in a structured, verifiable and trackable way. The EBP system will add to and enhance the development of the progressive digitisation of evidential

information on PHL. Developing the concept of EBP is a direct response to the call of Thomas Baker et al. in *Library linked data incubator group final report*.²⁶

There is a vast amount of prosopographical information hidden in documents that were produced in the nineteenth century (and earlier). Much of this information is messy and incomplete and it is the job of archivists to preserve that information in its current state. Making embedded prosopographical information digitally accessible is achievable with the adoption of current systems of information identification and data organisation. The skills and tools needed to achieve this are also readily available. A micro-study of typical prosopographical nineteenth-century data analysed in the P7 Case Studies will prove this. This thesis imagines how the human story might be more completely and more inclusively told if DH recognised that the scientific management of both the physical information of PHL (EBPI) and its representation as digital data (EBPD) were at the centre of the DH enterprise. In undertaking this enterprise and working under a greater humanities umbrella, DH could become the service provider essential to both the scientific and humanistic communities who need to study collectives of PHL.

This chapter opened with an observation: that there is a concern that take-up of the digital in the humanities is relatively low compared to other academic schools. Given that the humanities is a very broad academic area with a wide range of participant disciplines, no doubt there are many answers, each of which will be of more relevance to some schools rather than others. In answering the above questions, this thesis may be able to offer a view

²⁶ 'Linked Data could ultimately lead to new and better services to users as well as enabling implementers outside of libraries to create applications and services based on library data. It is too early to predict what new types of services may be developed for information discovery and use. Experimental services using library Linked Data should be undertaken in order to explore potential use cases and inform the direction of larger development efforts' (Baker et al. 2011, 11).

from a data science perspective, one that focuses on the universal relationship between an object of study (whatever it is) and its digital referent (whatever it is), because these underpin and overcome many of the challenges of changing technologies in rapidly developing disciplines.

1.8 Positionality Statement

My approach to academic work is primarily grounded in aspects of my Quaker faith and in particular the Quaker testament to equality.²⁷ In connection with the research presented here this has two manifestations:

- To value equally all past human lives and in action, to seek to preserve, reveal and cherish every instance of the evidences of past human lives. This means working to help correct the past practises that privilege research interest in only a few lives over the many (the famous, the wealthy, the exceptional). In the digital world we can (and I argue must) take the new opportunity to digitally reference and to some extent represent, the evidences of all past human lives. Today we have the technology and the expertise to rise to this challenge.
- To empower All Researchers, however skilled, however committed and however resourced, to be able to digitally find and work with the evidences of all past human lives. It is my concern that Independent Researchers living and working beyond the reach of institutional support are often unnecessarily overlooked, disadvantaged and therefore can be poorly served in the digital world.

²⁷ <https://www.quaker.org.uk/faith/our-values/equality>

With regard to the Project Seven case study my positionality led me to work with several experts each of whom contributed significantly to my work, and they are warmly acknowledged above. Project Seven considered data on around 3000 lives, all of which the study treated equally. The 3000 names included members of the aristocracy, leading scientists, and leading religious and military officers, but also many unremarked persons including for instance, the wives of ordinary citizens (sometimes identified in the data only as 'Mrs x'). No doubt some individuals present in the dataset played more active roles than others, both within the societies studied and as influencers in the political domain, - but current historical narratives based only on those few individuals prominent in the current historical record leave unrecognised and unaccounted others seemingly lost to the historical record. Thomas Hodgkin MD 1798-1866, was singled out here in a thesis case study that evaluated a published account of his lifelong activity in support of aborigines. The published account was carried out by the colonial historian, Professor Laidlaw. The thesis study had the express intent to show that there were many others unrecognised in the published account. In Project Seven all individuals were treated equally and given equal representation in the data. It was my objective in the study to show how at least 3000 people, over a considerable period of time, came together to support both financially and in presence, the work of relieving the plight of aborigines throughout the British colonies, and then also those amongst them who went on to support the emergence of the discipline of anthropology in Britain 1830-1870. Many of the 3000 are unnoticed by traditional historical narratives and therefore rendered invisible in published accounts. This thesis argues that the ability of the CEDA to endure and to be a significant thorn in the side of the British parliament's consciousness for so long must have depended on the support (in numbers at least) of most if not all of the 3000 group members. By including all 3000 members in the

Project Seven (detailed) visual analytical exercises I could be sure that I was trying not to allow the distortions of past historical practises to distort my study. Adopting this approach, and equally including all 3000 names, allowed Project Seven to demonstrate the main argument of this thesis, that all past human lives matter and that evidences of those lives should be digitally preserved if the recent developments in digital practice allow.

1.9 Chapter outlines

Chapter 1 (this chapter) has reflected on the disappointing level of take-up of digitisation in especially academic historical research. It has introduced the two-part nature of this thesis: (1) a practical case study underpinning (2) a critical assessment of the current state of infrastructures in DH as far as they are relevant for the development of EBP and the study of PHL. A taxonomy has been provided for the new and specialist terms used here.

In Chapter 2, 'Evidence Based Prosopography', beginning with the visions of Spina and Schöch, EBP and EBPD are fleshed out and explored in depth. The NAI system and the contribution that it will make to research into PHL using the Semantic Web are fully explained. Finally, how a selection of EBPD was chosen for the P7 project is described. The definition of EBP derived here will enable examination of the level of support for EBP currently available in DH.

In Chapter 3, 'EBP in the Digital Humanities', EBPD in DH (in the context and temporality of this thesis) and its relevance in current DH infrastructures are explored. It is discovered that high-level concepts in DH have failed to recognise the importance of EBPD (and the

importance of records themselves). This is clear in examining the work of leading DH thinkers, for example in both the 'Methodological Commons' (Unsworth 2000) and the 'Big Tent' conceptualisations (Terras 2016). This omission arose in the early history of DH. Finally, how EBP is currently supported in regional and local infrastructures is outlined. The examination of infrastructural support for EBP provides the context within which to examine the current affordances in GLAMS and the extent to which EBP is present in digitised finding aids.

Chapter 4, 'EBP in Galleries, Libraries, Archives, Museums and Special Collections (GLAMS)', takes a close look at digital archival Records in GLAMS, given that a lot of digitising effort has gone into the digitisation of archival Records and that some archivists are now using this metadata as a surrogate for records. A bewildering number of standards have been introduced into the sector over the last 40 years, forcing archives to commit considerable time and resources to their implementation. Not all archives have managed to keep up. How the evolution of standards has influenced the development of metadata systems in the US, the EU and the UK is examined to reveal the pressure archives have managed in conforming older finding aids to the new digital systems and the complexities and compromises made to integrate systems. It is discovered that digital record systems in GLAMS often provide unattributed notes on persons that are sometimes loosely related to the records that underpin those notes.. In this sense, GLAMS digital records are themselves secondary sources, in the meaning understood by historians. Given that researchers' main interest is in primary sources, GLAMS data can point the way to secondary sources and even provide a commentary on them, but they are not a substitute for EBP contained in all archival Records of PHL. This discussion of digital affordances in GLAMS enables a detailed

examination of recent digital projects that have sought to use archival digital catalogue data in research projects that are based in EBP.

Chapter 5, 'EBP in recent research projects', examines in detail six research affordances developed contemporaneously with this thesis. Their characteristics are compared to better understand how the regional differences explored in Chapter 3 can result in differences in affordances. Their utility as research affordances is considered and they are compared against each other using criteria developed in the P7 project. They were found to have a wide range of performance outcomes and one, the Golden Agents project, was found to be using EBP in a local project broadly similar to the EBP system proposed here. This discussion of recent projects that rely on EBP provides the context for an examination of my own P7, the practical exercise undertaken to support the theoretical critique of this thesis.

Chapter 6, 'EBP in Project Seven (P7)', introduces the P7 project, its definition, data requirements and characteristics. The project design and HDDT suite of technologies and the model built for the project are described. Issues in data management and how the project dealt with data problems are discussed. The P7 project is assessed as a standalone project undertaken by a Independent Researcher, unfunded and with minimal support. Finally, the P7 project is compared to the projects discussed in Chapter 5.

Chapter 7, 'Conclusion', begins with a reflection on both the learning and the experience gained from designing and building the P7 practical part of the work of this thesis discussed in Chapter 6, in combination with the analytical work set out in Chapters 3 to 5, and in particular the EBP and NAI system described in Chapter 2. The research questions set out in the current chapter are then returned to and answered. A response to the observation at the opening of this chapter, that take-up of the digital in those parts of the humanities that

deal with PHL is poor, is offered from a data science perspective. Recommendations are made and an impact statement sets out the advantages for the study of PHL if the EBP system recommendations made in this thesis were to be pursued.

Chapter 2 Evidence Based Prosopography (EBP)

2.1 Introduction

This chapter defines Evidence Based Prosopography (EBP) in the context of the Digital Humanities and explains its dual nature, where EBP is present as often messy information in the physical records of past human lives, and also in digital form in archival and genealogical records, as well as in the digital research affordances and projects that have EBP as an essential data component, such as the Project Seven undertaken by me for this thesis. This dual nature (physical / digital), together with the three broad areas of representation for digital EBP (genealogy, archives and research) frame the area of research for this thesis.

It is shown here that current genealogical practises and platforms already largely conform to good practise in EBP, albeit within strictly commercial enterprises. However, EBP in the archive considerably lags behind EBP in genealogical affordances, and researchers in turn, are largely unsupported in terms of relevant infrastructure. The independent National Authority Index system is then explained as a proposed solution, by seamlessly linking genealogical and archival records, together with the researchers who rely on those records.

First, EBP is placed within the wider context of the possible future development of DH. The challenges of dealing with the reality of messy data and the opportunities that arise if EBP can be systematised are explained in Section 2.2. Section 2.3 discusses EBPI, the incidences of EBP present in Records at archives and in genealogy, beginning with its root definition in prosopography. Section 2.4 examines in detail the problems of fixity and affixedness that

inevitably arise when EBPI is represented as EBPD. Section 2.5 explains that those problems arose in the P7 project and outlines how they were dealt with there. In response to the concerns identified in Sections 2.2–2.5, Section 2.6 explains how genealogy has overcome issues of fixity and affixedness and how attempts have been made in the recent past to build an authority index to help systematise the relationship between the multiple instances of EBPD found in genealogical data, and also to reveal the familial relationships hidden in the data with the advent of the Semantic Web. This is important because this thesis argues that systematisation of all instances of EBPD through authority indexing is central to extending infrastructure provision to cover research activities. The discussion shows how the few recent attempts at authority indexing in genealogy have failed to overcome the commercial drivers of genealogical affordance providers and their need to protect proprietary data. The solution proposed in this thesis is to place the authority index file at national level, separate from genealogical and archival affordances. Section 2.7 describes the EBP system and its component NAI-UID system in detail, setting out how it provides the research infrastructure necessary to address the considerable problems that arise when researching PHL. The chapter finally returns in Section 2.8 to the question posed above and concludes that the EBP and NAI system address common concerns in data integrity at the research level.

2.2 The DH context of EBP

Patrik Svensson sees DH, by its embrace of technology, expanding its reach beyond traditional boundaries within the academy and beyond the academy into wider society.²⁸

Salvatore Spina defines a powerful vision for the future of DH: the collection of all data on PHL into one great database organised through genetic structures.²⁹

These are two bold visions and what underpins them is expansionism. It is time, these authors argue, to imagine a future where DH is much bigger than it is today. EBP is a proposal that rises to the challenges of Svensson and Spina and allows DH to expand its interest and its reach. EBP as a system is ‘big data’ and requires a ‘big data’ approach if Spina’s and Svensson’s bold and complementary visions are to be achieved.

While it is beyond the scope of this thesis to address integrating genetics with big DH, it is recognised that genetics as an organising principle in the study of PHL is possible as a longer-term objective, and this acts as a reminder that any proposal made today for the immediate future of DH must also be open to the longer-term future. The integration of genealogy with archival practice in DH proposed here is open to the future. Moreover, this chapter will show that EBP is a credible and achievable next step in that journey towards the

²⁸ ‘Big DH describes a broadly defined, open, and challenging field that exists between humanities departments, disciplines, and epistemic traditions, between the humanities and other knowledge domains, and between the academy and the world outside. This position is driven by intellectual curiosity, technological imaginaries, historical sensibility, scholarly challenges, and a willingness to engage critically and technologically across issues, perspectives, and needs relevant to understanding and improving the human condition’ (Svensson 2016, X).

²⁹ ‘Our future lies in the archives and their heritage, through which we write history. But, if the historical archives represent the “databases” of the past, the “future of the past” lies in the greatest database that history has ever created: The genetic heritage, which is, on the one hand, the “a priori” structure of the performance of man’s action, and, on the other hand, a natural archive that lies inside every one of us. Written documents, atoms, and genes, nowadays, can enrich our historical research and give us the opportunity to understand the events and the men of the past.’ <https://www.timemachine.eu/ambassadors/salvatore-spina> (Accessed 20 September 2022).

future of DH. EBP is concerned with information on PHL contained in the instances of sources in the archive, in genealogical Records and other digitally accessible locations.³⁰

However, there is a dilemma at the heart of the expansionist technological visions of the future of DH: the current unsuitability of the data needed to fulfil that vision.³¹ To illustrate this dilemma, Christof Schöch provides a conceptual model on which Evidence Based Prosopographical Information (EBPI), as an example of the kind of disorganised data that will need to be ordered, could be mapped (Schöch 2013). Understanding this dilemma points the way ahead for EBP. Schöch expresses the digitisation dilemma as a journey we are on, away from messy data towards 'big smart' data, but we have not yet completed that journey. He argues that the DH journey begins with the first small and simple digital datasets made over thirty years ago, and leads in the future to big and smart datasets where big and smart data will be structured, organised and fully supported by metadata and appropriate infrastructures.

The problem with Evidence Based Prosopographical Data (EBPD) and Svensson's model is that EBPI, although it has the potential to be big data, is inherently and permanently messy information, and yet it must be preserved as it is, as messy information, because it is evidence of PHL. The dilemma is therefore that digital infrastructures are needed to overlay, but not replace, messy EBPI so that as EBPD it can be treated digitally as if it were smart.

³⁰ '[T]he term "data" applies not in the sense that information is quantitative, as it often is in the sciences and engineering, but rather in the sense that it is, like the quantitative information generated in scientific laboratories, primary source evidence for further investigation' (Waters 2023, 94).

³¹ 'In terms of the availability of information, we begin to see the emergence of a qualitative difference between digital history approaches to different historical periods. In mediaeval history there is a relatively small amount of data, much of which is textual. Although far too much for any individual to assimilate, it does mean that the corpus, while large, is approachable with digital tools. By contrast a 20th century historian faces a relative abundance of data, in many media, including audio and video. As well as posing problems of just how such a mass can be sifted, [] a great deal of material is still in copyright and may not only be difficult to access but also to analyse and reproduce. Historians of the 21st century will, given the exponential increase in digital data, find these difficulties much exacerbated' (Blaney et al. 2021, 11).

Schöch advocates two possible solutions to the huge task of cleaning up messy data: automatic cleaning (using algorithms) and crowdsourcing (using human interpretive skills) to achieve smart big data³² (marked 4 in Figure 2.1).

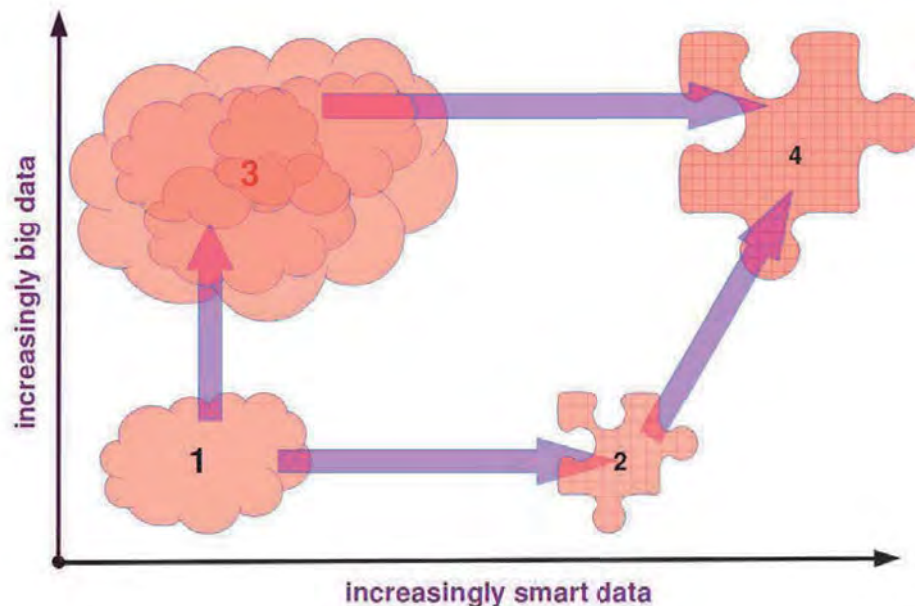


Figure 2.1 The story of smart and big data (Schöch 2013, 11)

In the case of EBP, it is likely that the system can become a crowdsourcing endeavour using large numbers of volunteers, working on little tasks and in small units to support archivists and researchers in the task of identifying and coding instances of EBP found in Records (see Section 5.5). The Golden Agents project successfully adopted this approach to coding project based EBP. The NAI system can also exploit crowdsourcing initiatives alongside those already adopted by genealogy platforms; such volunteer models in national index building are already in place and are working successfully in genealogy (see Section 2.6). It is

³² 'What we need is bigger smart data or smarter big data, and to create and use it, we need to make use of new methods. So, how can we enrich big data sufficiently to make more intelligent queries possible? How can we speed up the process of creating smart data so that we can produce larger volumes of it? Basically, there are two possible ways to do this: one is automatic annotation, the other is crowdsourcing. Automation refers to various heuristics of discovering implicit units, structures, patterns, and relations, and of making them explicit in the data. Crowdsourcing, on the other hand, relies on breaking down a large task into such small units that each of these little tasks can be performed in a distributed way by a large number of volunteers' (Schöch 2013, 10).

therefore possible to find an organising methodology for the digitisation of EBPI that will maximise its exploitation, one rooted in the simultaneous adoption of both of the solutions suggested by Schöch and Svensson, and one which can be rolled out across UK academia now, by exploiting both Semantic Web networking tools and crowdsourcing initiatives.

It is important to recognise that the overall process of making digital EBP benefits physical EBPI by helping to make EBPI easier to access, even remotely. At the same time, EBPI will retain its special status and the need for the preservation of EBPI will be enhanced and not diminished, because the digital copy will not replace the original. EBP will have as a core objective the preservation and good management of the relationship between originals and digital copies.

EBP will allow for later re-verification of datasets that were made before adoption of the NAI-UID system. The NAI-UID system will support provenance tracking from digital copy back to original source.³³ Adopting the NAI-UID system will help bring to light some of the complexities in the diffusion and reproduction of digital representations that lie beneath the Spina/Schöch and Svensson visions and offer a possible universal solution to messy data and degradation of information through digital dissemination. The rapid move to the Semantic Web, with its primary focus on relationships and relational network building (both personal and non-personal networks), invites more and better-connected relational data by design.

The EBP system includes the critical assessment of the limitations and vulnerabilities of digital representations of instances of EBPI in Records, which will grow in complexity over time as cultural distance grows between the digital and non-digital worlds. Lack of a

³³ 'Humanities computing, as a practice of knowledge representation, grapples with this realisation that its representations are surrogates in a very self-conscious way, more self-conscious, I would say than we generally are in the humanities when we "represent" the objects of our attention in essays, books, and lectures' (Unsworth 2002, 39).

national data authority system means that the digital representations of data (EBPD) risk becoming unmanageable as each instance of EBPD becomes more tenuously related to the physical Record it purports to represent. EBP critically analyses the advantages and limitations when and where EBPI and EBPD interact, and also where from time to time digital EBPD must be used as a substitute for physical EBPI, for instance when the physical information at the Record is lost, destroyed or cannot be located. EBP encourages the application of Semantic Web concepts to organise and structure data about PHL, for example by promoting universal authority indexing, the efficient use of persistent identifiers for all digital representations of EBPI found in sources, and the use of technologies such as Knowledge Graphs to help connect EBPD through networks.

Issues also arise at the data level itself, such as fixity and affixedness, and these too can be addressed by the NAI-UID component of the EBP system. They are discussed at length in Section 4.1, where these issues arise in GLAMS, and in Section 5.1, where they arise in researcher affordances.

2.3 Prosopographical information

Instances of EBPI are found in the Records at archival collections. This study recognises prosopography as it is defined in *A short manual to the art of prosopography* by Koenraad Verboven, Myriam Carlier and Jan Dumolyn (Verboven, Carlier, and Dumolyn 2007). The manual provides four distinct but related definitions from various contributors. The first is

that prosopography is a collected biography.³⁴ The second definition qualifies the extent to which prosopography is a collective biography by rejecting characteristics that describe personality and focusing on only material characteristics that describe persons.³⁵ The third definition asserts that prosopography is the study of common characteristics.³⁶ Finally, prosopography is defined as a database of common characteristics.³⁷

The essential characteristic that EBP takes from the discipline of prosopography is that both seek common material characteristics of PHL that are then collected in a database. Both prosopographical information and an EBP system comprise data about persons that can be ordered,³⁸ and EBP, like prosopographical information, consists of information about people, not their feelings (internal characteristics) or deeds (external actions). An important aspect of EBP is that instances of EBP found in sources are evidential and are referred to here as Evidence. Project Seven (Chapter 6) provides a practical example of prosopographical information as it is used in EBP. The P7 Case Studies are concerned with the instances of person names found in Records of varying types, recorded in different

³⁴ 'Prosopography is a collective biography, describing the external features of a population group that has something in common (profession, social origins, geographic origins, etc.). Starting from a questionnaire biographical data are collected about a well defined group of people. On the basis of these data answers may be found to historical questions (H. de Ridder-Symoens)', (N. BULST en PH. GENET, 'Introduction', in: IDEM (eds.), *Medieval Lives.*), (Verboven, Carlier, and Dumolyn 2007, 5).

³⁵ 'The prosopographical method consists of describing the material characteristics of a more or less homogeneous group of persons by collecting "the largest possible bundle of material elements allowing us to describe an individual, and those spiritual elements which would enable us to go from the person to the personality are excluded. Here lies the difference between prosopography and biography, though this does not mean that prosopography does not play an essential part in biography and vice-versa"' (Verboven, Carlier, and Dumolyn 2007, 5).

³⁶ 'Prosopography is the inquiry into common characteristics of a group of historical actors by means of a collective study of their lives (L. Stone)' (Verboven, Carlier, and Dumolyn 2007, 5).

³⁷ 'By "prosopography" we mean the database and the list of all persons from a specific milieu defined chronologically and geographically established preparatory to a processing of the prosopographical material from various historical angles, though some German historians would distinguish this second stage as "Historized Personenforschung" (N. Bulst)' (Verboven, Carlier, and Dumolyn 2007, 5).

³⁸ '[B]oth EAC and SNAC build on a long research tradition in the humanities of prosopography. Prosopographies are biographical dictionaries that identify groups of actors in their historical context' (Waters 2023, 93).

forms and at several different archives. In the P7 Case Studies each selected person name identified at archives resulted in the creation of a record in a single SQLite database called the ‘Centres for the Emergence of the Discipline of Anthropology in Britain’ (CEDA), and in the database the primary field is ‘person_name’ (see Table 2.1).

Royal Anthropological Institute – database of person names with attributes (members of founder societies)	1988 records copied to CEDA database
Quaker Committee on the Aborigines – names found on manuscripts at Quaker archives, London	16 records created in CEDA database
Members of the Aborigines Protection Society – read on microfilm of the society’s meetings held at RAI archives	1090 records created in CEDA database

Table 2.1 The CEDA database records (<https://www.w3.org/RDF>, Accessed 13/09/2023)

Compiling a project database comprising only EBPD on 3000 CEDA members was an exercise prone to human error and required frequent judgement calls. Complexities that commonly apply to sources include variations in name spellings, name abbreviations and the repetition of the same names that might or might not indicate two identical entries (John Hodgkin has a son also named John Hodgkin and Thomas Hodgkin has a nephew called Thomas Hodgkin, all of whom were subscribers to the APS and members of the CEDA database). Typing errors

and data transcription errors could occur because the data was manually extracted. Even if the data had been machine read, that would not necessarily guarantee greater accuracy of transfer.

2.4 Digital EBP

As the digitisation of information contained in Records continues at pace across academia, each archive chooses its own digitisation regime and information selection criteria using different cataloguing and storage systems, etc. This means that the number and variety of digital representations of physical Records propagate across the internet, from one archive and research hub to another and from the initial digital representation to multiple secondary digital representations. This variety and diffusion become another problem for EBP.

Digital representations can be singular, one digital record for one physical Record, but they can also be (and often are) representations of whole digital collections derived from multiple physical Records and captured in one digital file. Finding data boundaries across multiple digital instances requires judgement, as do finding and matching data across collections.

Digital representations made by academics are usually carefully managed and preserved by their owner, whereas unattributed copies of those representations found in research projects are often managed in much less disciplined regimes, and often outside of the scrutiny of the owner of the physical Record. There is a risk of information loss, loss of integrity and misrepresentation, as data moves from one digital representation to another.

Digital representations are frequently made for specific enterprises and so the focus at the time of digitisation can be much narrower than would be the case if the intent were to ensure that the entire Record be faithfully captured. For example, the obverse of a physical Record (document) may hold additional, valuable information, but that would mean making two digital images for a Record, when the digitising project may only permit one image per Record.

As digital representations multiply and spread across the web, the fact that they have been subjected to different data management, data handling methodologies and different processes means there is a need to be able to trace back to at least the primary digital representation and (hopefully) the physical Record. Assessments must be made about each data item each time a new research project gathers data to understand the provenance and critically assess the appropriateness of the digital representations selected for study.

Over time, the links back from digital representations to the physical Record are likely to be broken. Physical sources and primary digital representations can also be lost. Eventually it will be difficult to verify the authenticity or completeness of most digital representations. Secondary digital representations may also change from one digital representation to another, for example a set of digital images may be enhanced in a secondary collection.

As digital skills among the general population grow and academics become less comfortable working with original Records, skills in source criticism may decline and become an even more specialist skill than they are currently. The digital past risks becoming fractured from the physical past and so EBP is an essential skill that must be learned and developed now.

The table proposed by Koolen, Van Gorp and Van Ossenbruggendiscussed in Section 3.3 (Table 3,2) has been adapted here to show the complex relationship between Records and their digital representations (Table 2.2).

Record (physical) at archive	Digital representation
<p>■ Who created the text?</p>	<p>■ Who made the primary digital representation?</p> <p>■ Who made the secondary digital representation?</p>
<p>■ What kind of document is it?</p>	<p>■ What form is the primary digital representation (image, datafile, metadata, etc.)?</p> <p>■ What form is the secondary digital representation (image, datafile, metadata, etc.)?</p>
<p>■ Where was it made and distributed?</p>	<p>■ Where is the primary digital representation?</p> <p>■ Where is the secondary digital representation?</p>
<p>■ When was it made?</p>	<p>■ When was the primary digital representation made?</p>

	<ul style="list-style-type: none"> ■ When was the secondary digital representation made?
<ul style="list-style-type: none"> ■ Why was it made? 	<ul style="list-style-type: none"> ■ Why was the primary digital representation made? ■ Why was the secondary digital representation made?

Table 2.2 Records and their digital representations

2.5 An example of EBP from Project Seven

The P7 Case Studies relied on microfilm of nineteenth-century society reports, nineteenth-century manuscripts in archives, lists of nineteenth-century correspondence made by an archivist around 2010, the index to a recent publication and spreadsheets of genealogical data extracted by hand from genealogical software (Table 2.3). Care was taken to ensure the following:

- Data was traced back to physical Records and verified as true (within the scope of the project).
- The data selected was the best available for the purposes of the study.
- Data was comprehensive for the purposes of the study.
- Missing data and its implications for the study were assessed.
- Data was faithfully copied, error free, into the project dataset.

- Matching data from one Record to another was verified at the individual Record level.
- An appropriate authority index was created (in the absence of a national general authority index).
- Data was correctly and accurately described within the project.
- The project dataset complete with data documentation will be placed in an academic archive at completion

The Centres for the Emergence of the Discipline of Anthropology in Britain (CEDA) database	
Royal Anthropological Institute database of person names with attributes (members of founder societies)	1988 records with unverifiable attributes copied to Centres for the Emergence of the Discipline of Anthropology in Britain (CEDA) database
Quaker Committee on the Aborigines member names found on manuscripts at Quaker archives, London	16 records created in CEDA database (Names only)
Names of members of the Aborigines Protection Society read on microfilm of the society's meetings at RAI archives	1090 records created in CEDA database (Names Only)
<p>All of the above names additionally searched for in Wellcome Inst., Archives Hodgkin Papers series D, and in <i>Protecting the Empire's Humanity – Thomas Hodgkin and British Colonialism 1830 – 1870</i>. Prof. Zoe Laidlaw 2021, University of Melbourne. (looking for connections to Thomas Hodgkin).</p> <p>A Genealogist then found 600 Quakers amongst them and their family connections. (Ben Beck, Secretary, Quaker Family History Society).</p>	

Table 2.3 Project Seven data sources

Project Seven (see Chapter 6) located the founding document of the Aborigines Protection Society in the Wellcome Institute Hodgkin Family Papers Collection. The APS was founded at Ratcliffe Meeting House. Robert Bell was in the chair and Thomas Hodgkin moved that the APS be formed (three days earlier Hodgkin had joined a reformed Quaker Committee on the

Heathen, urged that it be renamed the Quaker Committee on the Aborigines, and redirected its purpose towards active political engagement). James Backhouse is recorded as having contributed to the meeting. The Wellcome Institute document contains several instances of EBP and these are referred to in the Project Seven Report, which is the subject of Chapter 6. For an example of this methodology in use in the Golden Agents Project, see Section 5.6. Here we show how this source, and its EBP information, should be managed under the NAI-UID system (see Figure 2.2):

- Project Seven would be a registered project - UID 123451783 at either at the sponsoring institute repository or independently at an independent data repository – UID 123456784.
- The archive where the Record is housed would have a UID 123456781 and the UID of the archival Record – UID 123456782.
- All of the person names marked on the Record would be referenced to the NAI-UID database, -NAI UID xxxx yyyy zzzz.
- Incidences of EBP marked on the Record are given a project UID – UIDs 123456783 - 6.

Items of EBP contained in incidences are also given a UID (marked in salmon).

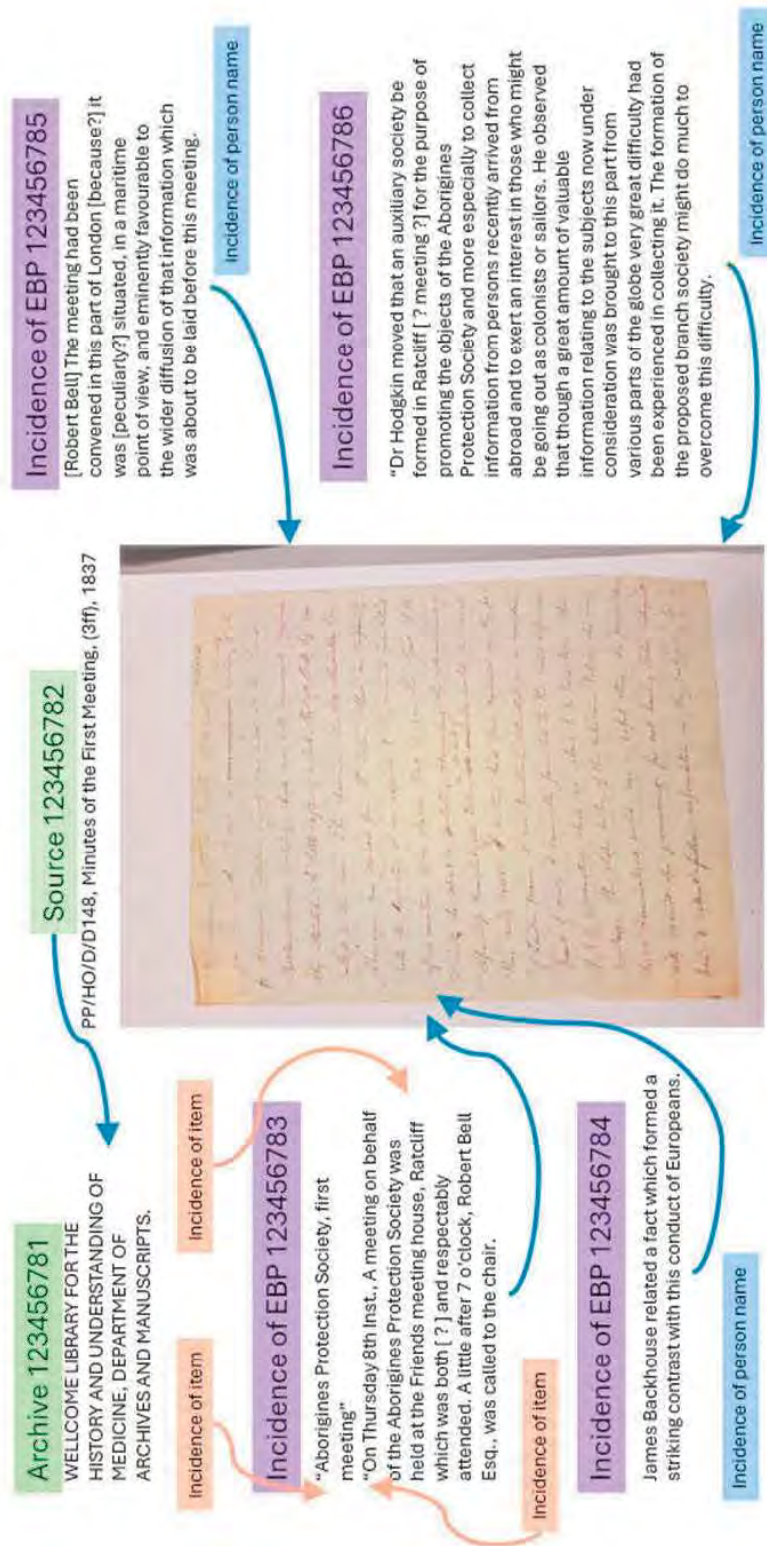


Figure 2.2 The NAI-UID system example

2.6 The role of EBP in genealogy

EBP brings together genealogy and the humanities through the deployment of digital technologies, seamlessly linking the world of physical sources (EBPI) to the world of representative data (EBPD) through the familial relationship structures of BMD records.³⁹ This universal structure was institutionalised in several countries by the state from early in the nineteenth century, and before then by the church. The historical structuring of national EBPI through familial relationships should be recognised as the basis for the building of an NAI, but this structuring is not without criticism.⁴⁰ It is not the intention here to reinforce paternalistic structures; however, any index of data on past lives must recognise the structures present in the data itself. J. de Groot explains how deep seated BMD structuring (the 'tree') is in family history in his 2015 'International Federation for Public History Plenary Address'.⁴¹ J. de Groot also warns that 'we' academics ignore 'amateur' historians at our

³⁹ 'The development of nineteenth-century institutions for recording family and social data coincides with historiographical approaches that sought to use the information collected by antiquaries and later the state to taxonomize the past. The association of the historian with the archive, and hence the perceived truth-telling qualities of the archive, are the basis of much early professional historiography. Although the archive collects information that is clearly not simply the materials of genealogy, it is the repository of social information that helps to constitute the nation-state' (de Groot 2015, 108).

⁴⁰ 'Genealogy is, fundamentally, something that cleaves to patriarchal models, to institutions of state, to the bureaucratic nation. It depends on particular kinds of knowledge, on training, on leisure, on access to archives, on hardware and software not available to much of the world. As a mode of historical knowing it might be considered profoundly Western, both in its models of family linearity as well as in its modes of investigation' (de Groot 2015, 106).

⁴¹ 'At their most basic level family history and genealogy involve the construction of datasets using particular types of evidence. The "tree" is the most common visualization tool. Family historians consult archives, visit collections, sift textual and visual evidence, collect private information, look at folk and local lore, consider monuments, visit churchyards, and look at surname registers. The types of archives consulted would form the basis for most historical investigation: church records; tax records; birth/marriage/death/burial information; census records; immigration records; ship passenger records; newspapers and journals (for obituaries mainly); wills; and military information of all kinds. Users have mastered most of the technologies associated with archival research: microfilms, microfiche, online databases, journals, card indexes, palaeography, codicology, and handling skills. They also collaborate and share information. The spirit of volunteerism in the family history community is strong, and users have created and shared many repositories of information and archives (as well as crowdsourced to aid projects)' (de Groot 2015, 110).

peril. The insertion of ‘amateur’ in quotation marks suggests that de Groot thinks academics consider themselves to be separate from such amateurs:

The exponential growth in genealogy is part of the first wave of the epistemological and historiographical shifts effected by the Internet. Around the world millions of people are choosing a particular type of historical knowledge. They are actively participating in historical enquiry. They are contributing to a store of historical knowledge and engaging with historical and historiographical debate. This is not a motif that is being imposed, it seems to me, and we ignore such ‘amateur’ history at our peril. (de Groot 2015, 125)

This view is shared by C D Hoeve.⁴²

A review undertaken in 1999 for the Max Planck Institute for Demographic Research by Natalia Gavrilova and Leonid Gavrilov shows the vast extent of genealogical information potentially available that was identified for just this one biological study.⁴³ All of the

⁴² ‘Despite its growing popularity, there is a noticeable absence of references to the inclusion of genealogy and family history studies within the field of digital humanities. New forms of inclusiveness, particularly in production-coding and cultural analysis, closely align genealogy and family history with the core tenets practiced among humanities computing and digital humanities. This paper aims to prove that genealogy as family history should be finally recognized within this cohort, as it can serve as a valuable and innovative partner for advocacy and technological advancement of the field’ (Hoeve 2018, 215).

⁴³ ‘The main cause that hampered many previous biodemographic studies of human longevity is the lack of appropriate data. At the same time, many existing data resources (millions of genealogical records) are under-utilized, because their very existence is not widely known, let alone the quality and scientific value of these data sets are not yet validated. The purpose of this work is to review the data resources that could be used in familial studies of human longevity’ (Gavrilova and Gavrilov 1999, Abstract).

‘The data collected for the study runs for some 35 pages and is collected into 5 sections. The study was published in 1999 and therefore the data review is 25 years old. This review consists of the following 5 sections: Section 2 Data resources developed for biodemographic studies of longevity. Section 3 Databases created for the studies in historical demography. Section 4 Data resources for long-lived persons and their families. Section 5 Computerized genealogical data (products available on the international market). Section 6 Published genealogical and family history data (that could be recommended for their use after computerization)’ (Gavrilova and Gavrilov 1999, 3).

‘Information presented in this review demonstrates that an enormous amount of data on familial longevity is available now for researchers and even more data became available since our first report on this topic. Millions of genealogical records are already computerized and could be potentially used for the study of the familial aggregation of human longevity. Most of these genealogies are a product of family reconstitution, carried out both by professional genealogists and by family members wishing to trace their ancestry back to the founder who brought their surname to America or even to their European family roots. The compilers of genealogies aided this time consuming task with the many different sources: genealogical libraries, LDS (Mormon) church family history centers, genealogical search engines available on Internet, computer CDs with census, marriage, land, probate records and many other resources for genealogical research. The potential of the existing data resources is understated and the data resources are underutilized. We hope that this review of data resources will stimulate further large-scale studies on the familial clustering of human longevity. (Gavrilova and Gavrilov 1999, 7).

information discussed by Gavrilova and Gavrilov is potential EBP. Once the EBP approach is widely adopted, DH will begin to open up future research opportunities across many academic fields, and even beyond the humanities themselves.

The skill sets and tools that a practised genealogist brings to the study of PHL are essential in the study of EBP. Although generally the relationship between genealogy and academia is still problematic (for both sides),⁴⁴ the pull of EBP (if it is to be championed by DH) should be sufficient to bridge the ‘trust’ gap between them, and thus progressively bring them together.⁴⁵ Joint working in academic disciplines and genealogy is growing (see Section 5.1). Formal accreditation through the Board for Certification of Genealogists⁴⁶ and award-winning historical/genealogical projects such as ‘The Valley of the Shadow’ help to make the gap between genealogy and academia more bridgeable.⁴⁷

⁴⁴ ‘Notable in this community is the lack of participation of archivists and librarians although they are enablers of the community. The interviewees did not place them centrally either in educating genealogists about records and the search process, or in the creation and disposition of family archives. If communities are defined by both participants and nonparticipants, this lack of participation is an important consideration for the archival community. Genealogists are supportive of archival activities; they do not, however, rely heavily on archivists for education, either about searching for records or about preserving family records’ (Yakel and Torres 2007, 111).

⁴⁵ ‘The shelves of genealogical and local historical societies are filled with histories of families whose prominence is generally confined to the locality, written by people still less well known. Most of these are not much more than padded genealogies and are not likely to be useful to the historian. However, the bare genealogical record—births, deaths, lines of descent—can be helpful in the study of family mobility and “in the technique of family reconstruction,” which is one of the aims of historical demography in studying the early American family ... Occasional papers urging cooperation by genealogists, historians and social scientists have gone for the most part unheeded’ (O’Hare 2002, pages unnumbered).

⁴⁶ <https://bcgcertification.org> (Accessed 2 November 2023).

⁴⁷ ‘The Valley of the Shadow: Two Communities in the American Civil War, co-authored by Edward L. Ayers and Anne S. Rubin, is an “invented archive” or cross-repository collection drawn together specifically to create an online resource. Valley of the Shadow takes two communities, one Northern and one Southern, through the American Civil War via an archive of sources: newspapers, letters, diaries, photographs, maps, church records, population census, agricultural census, and military records. As the site’s introductory text states, “Students can explore every dimension of the conflict and write their own histories, reconstructing the life stories of women, African Americans, farmers, politicians, soldiers, and families.” The prize-winning site was the focus of a *New York Times* article entitled “An Historian presents the Civil War, Online and Unfiltered by Historians” (June 29, 2000), and it is designed to operate as a do-it-yourself history kit, allowing users to track ordinary individuals from diary entries to newspaper articles to census records, without the mediation or structure imposed by an historian. The process encourages amateur research, and it creates the same sense of uneasiness in academicians (per Gary J. Kornblith’s review of the site in the *Journal of American History*): “in

Genealogy is popular among many people outside academia and it has a long history.⁴⁸ The popular digital platforms used for genealogical study today are more mature than the set of digital tools available to historians.⁴⁹ Genealogical datasets (already systematically organised and digitally searchable) are immensely valuable in EBPD, and its global and mass take-up means that it has had more time in use and therefore its faults were ironed out a long time ago.⁵⁰

2.6.1 Building a National Authority Index

A vast amount of EBPI from the mid-eighteenth century onwards could today be digitally structured and organised to form an NAI using UUIDs that record instances of person names found in national records of BMD data. These have been systematically collected at state

practice there is a thin line between destabilizing received narratives and promoting a nihilistic view that the historical record is so fragmented and complex that it makes no sense at all.” (O’Hare 2002, pages unnumbered).

⁴⁸ The beginnings of genealogy as a systematic and rigorous pursuit go back at least into the early years of the nineteenth century, for instance see ‘John Farmer and the making of American genealogy’ (Weil 2007).

⁴⁹ ‘Genealogy – From the Open Directory Project, this site indexes over 6,600 genealogy sites on the Internet. They are categorized by subject and geographic region. Links to foreign-language sites are also included. Genealogy.com – Offers tips on starting genealogical research, web links, and a 470 million-name searchable database. GenealogyToolbox.com – Resources on this site include a list of over 70,000 links to genealogical resources on the Internet, news and articles on family history research, a guide to genealogical software, and a site to register your genealogy home page.’ <https://www.archives.gov/research/genealogy/other-websites/database-links> (Accessed 1 November 2023).

⁵⁰ ‘For centuries, genealogy has been a model for historical investigation, often associated with antiquarianism and dynasty. It is a practice connected to heraldry, marriage negotiation, pedigree, and the organization of family. Colleges of arms still exist from their establishment in the early modern period (when inheritance, legitimacy, and the supporting family trees were so important they led to conflict). Laws associated with inheritance and primogeniture depend on linear models of familial process over time, visualized in the illustrations of heraldry and genealogy. Certainly it was the case in early modern and medieval genealogy that dynastic and matrimonial political expedience ensured that the private was public. Particularly in the West most medieval and early modern legal systems enshrined a basic patriarchal principle of male inheritance and surname organization. As a consequence the bulk of most national archives (outside of government documents) are made up of information relating to land, property, and family. Furthermore, the institutions of statehood and church have for centuries sought to understand, audit, and record the family relationships of their subjects, members, or citizens through censuses, the organization of key familial data (births/marriages/deaths)’ (de Groot 2015, 107).

level in Britain from July 1837.⁵¹ Earlier BMD data, most frequently found in the form of Parish Registers, is also structured, and it too can be organised using BMD. Census data which records household (and largely familial) data is also digitally available and that can help identify persons in BMD records.⁵² Helpfully, genealogical data is largely universally structured and where it is collected at state level it is also almost universally digitised. The universal and systematic structuring of the finding, organising and managing of EBPd in the study of family history, using digital genealogical affordances, is a major part of DH, and yet genealogy is today perhaps a poor relation to academia.⁵³

BMD records can be used to construct person name indexes, although because the genealogical affordances are proprietary each affordance provider will have its own commercially valuable index.⁵⁴ Adam Hjorthén, in 'An ocean of information: labour, commodification, and the culture of indexes in modern transatlantic genealogy', recognises the evolving and adaptive nature of indexes in genealogical affordance which this thesis adopts in the recommended development of the NAI system.⁵⁵ At least one attempt has

⁵¹ <https://www.gov.uk/research-family-history> (Accessed 7 March 2025).

⁵² 'The mid-nineteenth century was a period when the almost indiscriminate collection of statistics had become a mania, and the census can be seen as part of this movement to reveal the "state of the nation". The belief that certain laws, which were discoverable by empirical research, underlay creation was a very powerful strand in the intellectual make-up of the period' (Higgs et al. 2021, 7).

⁵³ Academia has long held genealogy at arm's length, as de Groot discusses in 'On genealogy' for the International Federation for Public History Plenary Address 2015: 'Genealogy as a model of historical knowing or an epistemology is largely untheorized by public historians. At present there is little historiography and no understanding of its qualities and impacts. This is surprising given the importance of such models in past centuries. At present it is even stranger, given the massive numbers of people investigating their past in the field that has become increasingly known as family history. Estimates of those involved in such activity regularly suggest that it is one of the biggest participatory activities on the planet' (de Groot 2015, 103). As H. Daniel Wagner notes in 'Genealogy as an academic discipline', 'no genealogy textbook currently exists and genealogy is not taught anywhere in a formal, academic setting', (Wagner 2006, 8).

⁵⁴ 'Indexes are used to gather, classify, organise, transform, store, retrieve, and communicate information. They come in different forms and contain different sets of data, but a common feature of genealogical indexes is that they all cross-reference information of value to the researcher, including names, birth dates, and birthplaces' (Hjorthén 2022, 190).

⁵⁵ 'Like genealogy in general, the history of indexes is shaped by both media and labour. Indexes are an example of a "knowledge infrastructure", created from hybrid processes of the "technical" and the "social".

been made to develop an interchangeable ontology for GEDCOM data, though this would be problematic in a competitive commercial environment.⁵⁶ The NAI system proposed here overcomes commercial sensitivities over proprietary data by locating the NAI within a national non-commercial entity (such as the General Register Office, GRO).

The structural organisation of familial relationships in all genealogical affordances is the GEDCOM file metadata protocol developed by the Church of Jesus Christ of Latter-day Saints (LDS) in 1985, but the disciplinisation of family history dates from at least 1974.⁵⁷ The GEDCOM protocol is today considered inflexible and efforts are being made to update/replace it with a more efficient and flexible system, though the basic structure of tagged familial relationships is expected to remain.⁵⁸ Attempts have been made to establish an interchangeable ontology to facilitate the exchange of data between commercial providers, although this ontology would also face proprietary issues. There are several

Rather than being formed as fixed and coherent entities, they are adaptive and continuously changing' (Hjorthén 2022, 190).

⁵⁶ 'The Semantic Web is very suitable for publishing genealogical data in an open and extensible way. In this paper a first attempt is presented for a Genealogical Ontology, that can start the discussion for a standardized ontology, that improves the exchange of genealogical data and facilitates automatic processing. With such an ontology, standard software tools can be used to encode integrity checks on the data and perform intelligent processing' (Zandhuis 2005, 8).

⁵⁷ 'The Federation of Family History Societies (FFHS) was established in the UK as an educational charity in 1974, and now has some 180 societies globally; the Federation of Genealogical Societies (FGS) was founded in the United States in 1976 and the African American Family History Association (AAFHA) was founded in 1977. *Family Tree* magazine was published first in the UK in 1984' (de Groot 2015, 109).

⁵⁸ 'In 1985 the Family History Department of the LDS released a propriety open-source, freely available coding file type known as GEDCOM (GEnealogical Data COMMunication). Such files contain plain text linked by meta data. GEDCOM files record information according to a particular protocol. They arrange information into easily tagged elements. Data is recorded, formatted, and then may be shared. GEDCOM records are viewed with a particular type of software that interprets the coding. Genealogical information is therefore arranged in a particularly inflexible way. Nonlinear or nonnormative interpersonal relationships are unrecordable, although most genealogy software will allow the user to add in such information manually to records. GEDCOM allows the sharing of information with other genealogical users and hence has become the basic mode of encoding. Nearly all genealogical databases, software, and websites run GEDCOM as their basic information template. The general structure for encoding information about the past and then circulating it is derived from the Mormon model. Hence the literal model of the past, the evidentiary unit, is predicated upon the collecting of information relevant to a particular religious identity. The individual in the past is reduced to a set of database tropes and tags. They are part of a broader taxonomy that is concerned with ordering' (de Groot 2015, 115).

major affordances in the sector, all of which use the GEDCOM protocol, with Ancestry.com perhaps being the largest.⁵⁹

2.6.2 An early LDS project using a prototype genealogical index

In 2009–2011 Douglas J. Kennard, William B. Lund and Bryan S. Morse completed a project to discover Historical Social Networks (HSN) embedded in two datasets by relating them to a proxy genealogical index. The project was called ‘Improving historical research by linking digital library information to a global genealogical database’ and it was considered the first study of its kind.⁶⁰ The team analysed the LDS Pioneer database and the LDS Migration Database, extracting EBP on the persons recorded.⁶¹ The extracted records were then compared to the FamilySearch Index.⁶² An illustration of the project appears as Figure 2.3.

⁵⁹ ‘Ancestry.com, for instance, generally considered the biggest provider, claims the following worldwide:

- 14 billion family history records
- 60 million member trees
- 6 billion profiles uploaded to those trees by Ancestry users
- 2.7 million worldwide subscribers
- 1,000 employees worldwide
- 200 million photographs, scanned documents and stories uploaded’ (de Groot 2015, 115).

⁶⁰ ‘We are not aware of other systems that use an existing genealogical database as an authority control framework for written DL [Digital Library] artifacts’ (Kennard, Lund, and Morse 2009, 256).

⁶¹ ‘We use two online databases of historical information and documents for our research into automatically discovering HSNs and linking the people in the networks to the FamilySearch genealogical database. The first is the Pioneer data base and the second is the Migration database. We have crawled the website of each database to record the roster of each pioneer company or ship voyage, record the information associated with each person on each roster, and store the URLs from where the information is retrieved. We also store the URL of trail or voyage excerpts that can be used to detect HSNs’ (Kennard, Kent, and Barrett 2011, 46).

⁶² ‘To link people in the Pioneer and Migration databases to the FamilySearch genealogical database, we use the search function of the Internet API provided by FamilySearch to find the people and then store the FamilySearch PersonIDs associated with them. The API search function returns a list of query results ranked by how well given query parameters match the people in the database. Each result is also classified as either a “close” or “partial” match, depending on whether the match score meets some minimum threshold. The parameters that may be used for a search query include the person's name, gender, event dates and places (for birth, marriage, and death), as well as the name and event dates and places for the person's father and mother’ (Kennard, Kent, and Barrett 2011, 47).

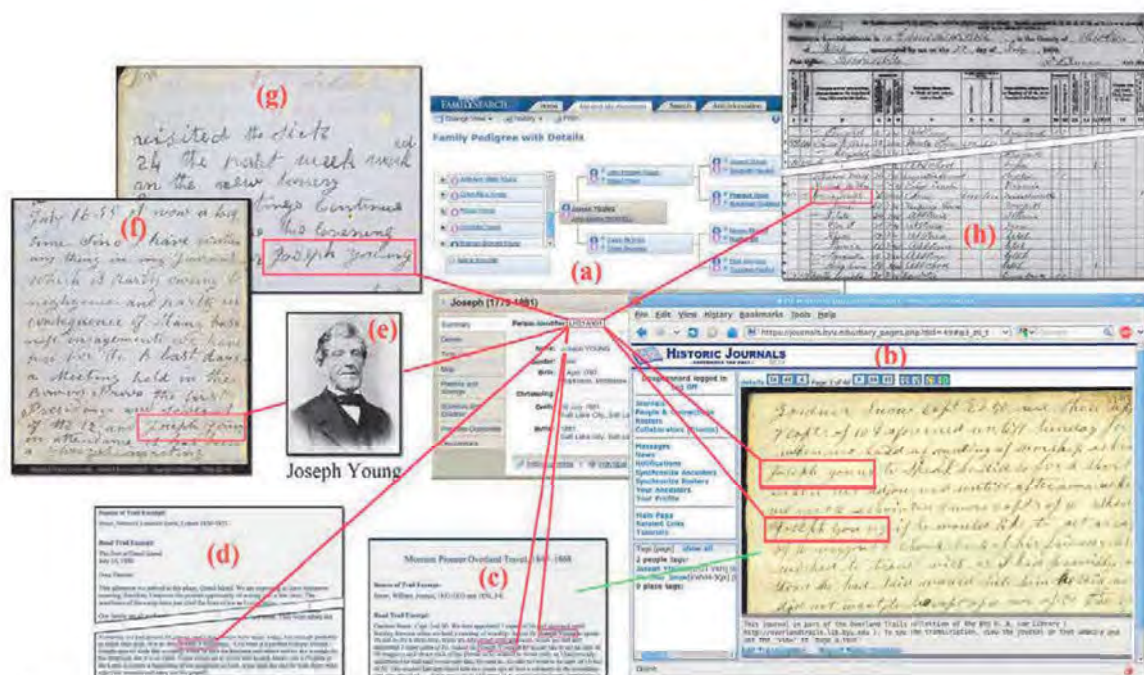


Figure 8: Links generated from a Historical Social Network discovered for American pioneer, Joseph Young: (a) Interface to FamilySearch database containing PersonID (red box) and links to family members. (b) Interface for Historic Journals website with tagged references to Joseph Young. (c) Searchable transcription of journal shown in (d) with links to tags. (d) Another trail excerpt that refers to Joseph Young. (e) Photograph of Joseph Young. (f) and (g) Other journals that reference Joseph Young, with links to tagged references. (h) 1870 Census record containing the name of Joseph Young and family members.

Figure 2.3 Finding EBP in the HSN Project (Kennard, Kent, and Barrett 2011, 49)

The project prefigures the work of the Golden Agents project discussed in Section 5.6, and both projects indicate the application of EBP in the arena of family history and the suitability of genealogical affordances as contributors to the NAI system.

2.6.3 Project Seven

Regarding the P7 Case Studies in Chapter 6, this researcher worked with an independent and reputable genealogist (a member and officer of the Quaker Family History Society),⁶³ a collaboration which was successful. There were minor technical challenges with data

⁶³ <https://newtrial.qfhs.co.uk> (Accessed 20 March 2025).

integration between the genealogical platform used by the genealogist and the HDDT database, but these were quickly overcome. The genealogist also from time to time questioned the accuracy of some of the data donated by other institutions (and the author's own research), offering suggestions that improved the quality of the data. The result was an opportunity to extend the scope of the P7 Case Studies analysis through the production of vivid visualisations of the genealogical networks of the 600 Quakers among the 3000 individuals in the main group, and over forty years of their politically/scientifically active lives (1830–1870).

There may still be some academic reluctance to accept a role for genealogy in the mapping of the social connectivity of nineteenth-century societies, even when working alongside historians and research software engineers (RSEs), but this thesis argues that academics should make efforts to overcome these anxieties.⁶⁴ The value gained from welcoming genealogy into the fold, and being able to study individuals, families and social groups in their entirety as EBPd, is immense.⁶⁵ In the context of EBPd, genealogy and genealogists are perhaps the lead data providers, and if DH is indeed a 'Big Tent' (Svensson 2016) then welcoming genealogists inside the tent should be considered highly desirable. The Imperial

⁶⁴ '[G]enealogy and family history are undergoing a huge technological shift. Historians and cultural critics are still trying to conceptualize the ways in which the paradigm shifts of the Internet are changing the ways that we conceptualize the past and understand our relationship to it... Genealogists demonstrate varying levels of competency, professionalism, and expertise. GENUKI's "Getting Started in Genealogy and Family History" advice page counsels that, "there are accepted standards for doing genealogy 'properly'—standards that we recommend you seek to learn and emulate"' (de Groot 2015, 126-127).

⁶⁵ 'Genealogy as a model of historical knowing or an epistemology is largely untheorized by public historians. At present there is little historiography and no understanding of its qualities and impacts. This is surprising given the importance of such models in past centuries. At present it is even stranger, given the massive numbers of people investigating their past in the field that has become increasingly known as family history. Estimates of those involved in such activity regularly suggest that it is one of the biggest participatory activities on the planet' (de Groot 2015, 103).

War Museum's 'Lives of the First World War' project illustrates how genealogists and archivists can work together, supported by crowdsourcing.⁶⁶

2.7 The proposed NAI system

There appears to be a provenance or authority issue with digital data where digital data is meant to be a faithful digital representation of an archival or other Record of a PHL. Such provenance issues are common in archival systems.⁶⁷ This provenance or authority concern arises after the first digital record is made of the name (EBPD) and then later that digital record of the name is transcribed into a database, and then perhaps copied from that database to another and so on, with each transcriber trusting that the data has been 'faithfully copied' either manually or electronically⁶⁸ through the entire process of digital dissemination. It would be wise to recognise the difference between accredited EBPD and unaccredited EBPD. There is a serious risk of loss of provenance in the process of data veracity if digital dissemination remains relatively undisciplined as it is today, which could

⁶⁶ 'IWM's Lives of the First World War tells the stories of individuals from across Britain and the Commonwealth who served in uniform and worked on the home front. This innovative digital project ran from 12 May 2014 to 19 March 2019. From individuals and families to communities and organisations, more than 160,000 people collaborated to piece together the lives of people who experienced the conflict, through sharing anecdotes and digitising material that has been hidden away in attics until now.' <https://livesofthefirstworldwar.iwm.org.uk/about> (Accessed 10 October 2023).

⁶⁷ 'The main difficulty is that sources often exist in multiple versions, either as a feature of their original production or because they have been copied and disseminated over time. In their cataloging, researchers must trace and account for the provenance and reliability of the versions they are using. Because digitization produces yet additional versions in a medium where it is easy for copies to proliferate, cataloging them is both more complicated and essential' (Waters 2023, 91).

⁶⁸ The term 'faithfully copied' is used here to describe the faithful copying of data from one repository to another, without alteration. The new representation of the data must be an exact copy of the primary digital form.

result in a future world of mass digital data provision on PHL that has questionable provenance and authority.

A digital representation of EBPI embedded in a physical Record can only be an authorised EBPd representation if it points unmediated to an original instance of EBPI in a physical Record in the archive, in which case, this thesis argues, the archival copy EBPd ought to bear a mark of authorisation made by the relevant archivist. Then in turn, the research dataset EBPd instance needs its own mark of authorisation made by the relevant researcher. It must be recognised that if DH data on PHL develops without systematisation, then future researcher datasets may point to a bewildering number of disseminated digital records, with no viable means of tracing the path from the digital instance in the research project to the original approved EBPd digital record made by the EBPI host archive.⁶⁹

A disciplined system for the management of digital representations of EBPI embedded in physical sources is the first step in addressing data integrity concerns related to infrastructure building in EBP. The organising concept at the heart of the EBP system is the National Authority Index (NAI) system, which in the UK could be based on the GRO index of births (Figure 2.4). The top level of the EBP system is the NAI UID. This has at least three basic elements of EBP which tie it to a PHL:

- The person's name (in any variation and not always clearly legible).

⁶⁹ The terms 'primary' and 'secondary' used here must not be confused with the uses of these terms to categorise sources in history. In history a secondary source refers to a primary source. In DH practice secondary means a duplicate copy, in this case of an instance of EBPd taken from another dataset which in its turn may reference an earlier original instance of the EBPI archive. (In practice it could be many steps back before a primary Record is found.) A digital secondary source therefore is a copy of a copy. This concern is managed by the adoption of the NAI infrastructure with its system of UIDs for all digital representations of EBPd.

The NAI can now be built to include EBPI from July 1837 onwards using national BMD certificates, and prior to that EBPd taken from genealogical BMD records in local and ecclesiastical records. Both sources of EBPI are offered today as publicly available open-record EBPd by the GRO (see Figure 2.5).⁷⁰

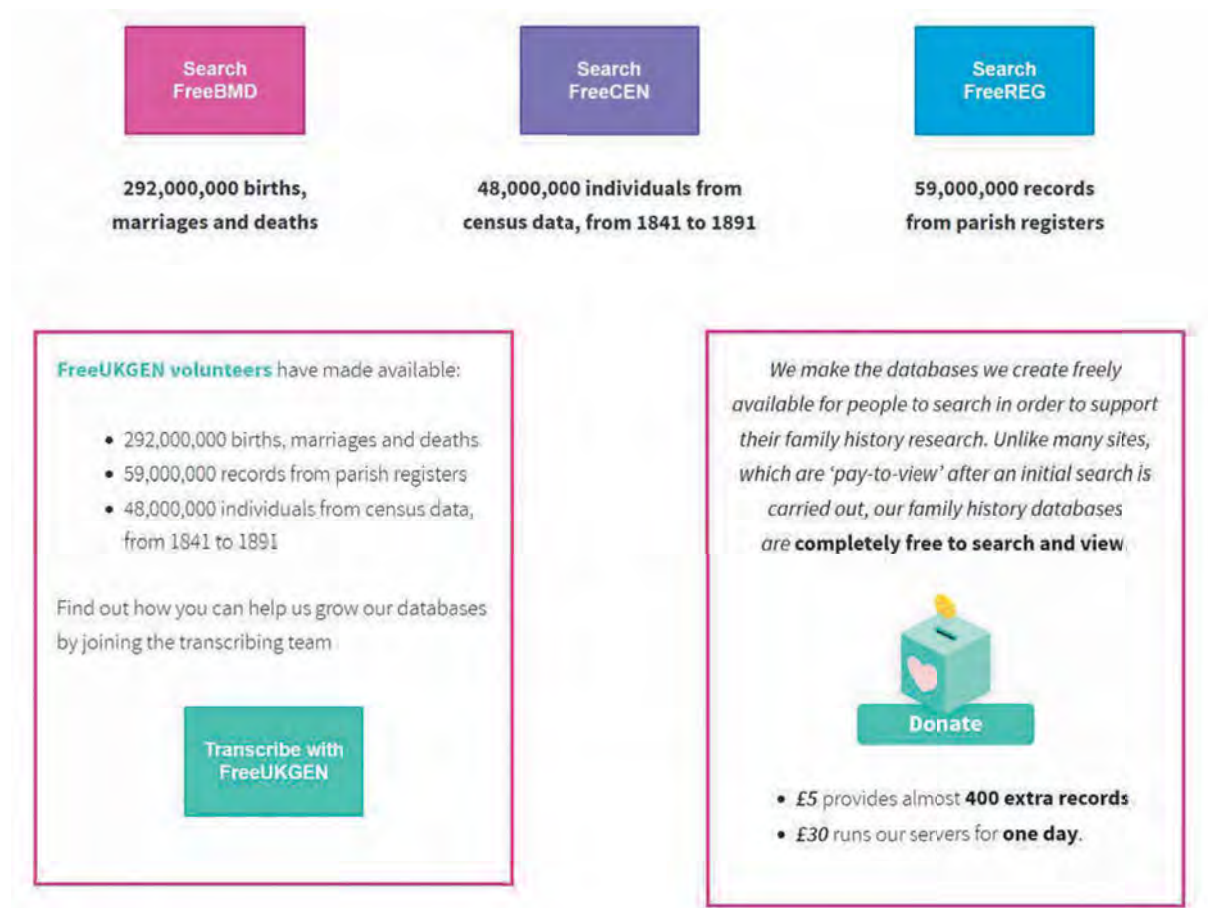



Figure 2.5 Free UK Genealogy CIO front page

⁷⁰ 'Free UK Genealogy CIO began in 1998, building a digital record of all registered births, marriages and deaths in England and Wales, and as of 25 Apr 2024 volunteers have created from digitised registers 294,092,936 distinct records (394,226,516 total records). Free UK Genealogy provides free, online access to family history records. We work with a team of dedicated volunteers to create high-quality transcriptions of public records from governmental sources, parish churches, and other trusted institutions. We believe that Open Data and Open-source are key to making and keeping public records accessible to all.' <https://www.freebmd.org.uk> (Accessed 27 May 2024).

The NAI-UID system can be extended back in time as far as BMD records allow (these records are commonly used for this purpose by genealogists).⁷¹ The NAI-UID will be a publicly available open-source affordance, although there are limitations: in the UK, for instance, it cannot include records less than 100 years old because it must be fully compliant with data protection obligations.⁷²

The person name in the NAI-UID will be the person name as it appears on the birth certificate and it usually has four locating attributes: mother, father, date, location (Figure 2.6).⁷³

Autocomplete



National Authority Index	Source UUID	Location
Thomas Hodgkin + DoB + M + F + Loc.	1234567890	GRO-UK

GRO birth record

Figure 2.6 The proposed National Authority Index

⁷¹ ‘Genealogists are an understudied group within the archival and library communities, although genealogy has been the focus of research in other fields, such as sociology. Sociologists examine genealogy as a cultural phenomenon, focusing on motivations for this activity and its underlying meaning to individuals or social groups. The archival and LIS [library and information science] literatures largely conceptualize genealogical research from a managerial perspective. One article by Wendy Duff and Catherine A. Johnson and another by Elizabeth Yakel articulate a different perspective, unique because they view the archives from the genealogists’ perspective. Duff and Johnson found that genealogists’ information-seeking patterns were as likely to work around existing archival systems as to use them and that genealogists’ search processes relied more heavily on their own social networks than on professional archivists’ (Yakel and Torres 2007, Abstract).

⁷² <https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted> (Accessed 27 May 2024).

⁷³ This representation is known as a ‘triple’ and it is the basic unit of the NAI system. All instances of past lives can be referenced in ‘triples’.

2.7.2 The Archival Authority Index

The second step is to extend the NAI-UID system to archives and collections,⁷⁴ through the adoption of a system of UIs for all instances of person names appearing in Records in the archive (see Figure 2.7). Almost all items held at archives and collections include recordings of person names. While allocating UIs to each instance of EBPI recorded in items held in the archive is a new application of Linked Open Data (LOD),⁷⁵ that archives should use existing authority systems to exploit LOD is not new, as Ioannis Papadakis, Konstantinos Kyprianos and Michalis Stefanidakis discuss at length.⁷⁶

⁷⁴ 'The term "archives" can be used in three different ways. Archives are: 1) the records, regardless of format, created or received by a person or organization during the conduct of affairs and preserved because they contain information of continuing value. Within an organization, the term "archives" refers specifically to the non-current records of the organization or institution, those records that are no longer required by employees to do their daily jobs, that document the growth, activities, and accomplishments of the organization; 2) the building or part of a building where archival materials are located (archival repository); 3) the agency or program responsible for selecting, acquiring, preserving, and making available archival materials' (Morris 2009).

⁷⁵ 'The focus of Linked Data on the graph as a whole rather than on bounded sets of data defined by their source has led many to value connections among data sets and become less invested in maintaining distinctions among them. The Semantic Web's open-world assumption posits that a lack of information must lead to an uncertain conclusion (a closed-world assumption, on the other hand, offers more certainty but relies on all relevant information already being known). This stance has led to a desire to more heavily link data from multiple sources, building value and new knowledge from these connections. Metadata creation in this context becomes less about filling in fields on data-entry forms than about making links between pre-existing things. The future may very well continue this trend towards deeper connections and larger knowledge graphs' (Riley 2017, 40).

⁷⁶ 'Traditionally, libraries provide access to collections via the employment of Online Public Access Catalogs (OPACs). The OPAC is a fundamental component of an Integrated Library System (ILS) since it facilitates access for the average user to information (both bibliographic and authority data) stored in Machine-Readable Cataloging (MARC) format. At the beginning, the main purpose of an OPAC was to aid users in locating books on the shelves and/or linking books that share a common aspect (e.g. subject). Along these lines, library professionals throughout the years have collected valuable and high quality, authoritative information that can be utilized beyond the scope of the library OPAC. This paper focuses on authority data and argues that such data should be made publicly available in a widely acceptable, machine-understandable format. Linked data technologies provide the means to render authority data within libraries part of the so-called Web of Data (WoD). The WoD refers to a vast amount of data on the web available in a standard, machine-readable format, which can be reached, linked and managed by adequate semantic web tools (Bizer, et al. , 2008). To meet this goal, traditional MARC-based authority records should be enriched with linked data-specific information (i.e. linked open data (lod) URIs)' (Papadakis, Kyprianos, and Stefanidakis 2015, 1). 'The advent of the semantic web, and the linked data movement in particular, provide the opportunity to promote access to authority records in a standardized manner. For this purpose, authority records within OPACs need to be updated with linked data-specific information' (Papadakis, Kyprianos and Stefanidakis 2015, 11).

Archive Authority Index	Identifiers	Location
Dr Hodgkin	2345678901	Physical at TNA
Thomas Hodgkin NAI	1234567890	GRO - UK
Embedded in Document	Dublin Core UUID	Physical at TNA

Links to Google Scholar

Archive record with UUID made for each incidence of a person name attaching it to NAI. New, can be made by AI

Every incidence of a person name at archive is allocated a UUID, then linked to the NAI UUID and Dublin Core UUID (for the document the name appears in).
Every item at archive has a physical location

3

Figure 2.7 The proposed Archival Authority Index

The Archival Authority Index record will include a UID for the incidence of the person name found in a Record at that archive. The person name is entered into the AAI in the form in which it is found in the Record (e.g. Dr Hodgkin). The freedom for each archivist, when compiling the individual collection's AAI, to choose each NAI-UID allocation is an essential characteristic of the NAI-UID system. For example, although the NAI-UID entry in Figure 2.7 is for 'Thomas Hodgkin', the archive might be a collection of Quaker Records and in the Quaker archive this person might be habitually referred to as Dr Hodgkin, so this would be

'MARC suggests a data format that is employed to exchange, use and interpret bibliographic and authority information between libraries, thus enhancing interoperability between them. It employs a system of numbers, letters, and symbols to annotate information' (Papadakis, Kyprianos and Stefanidakis 2015, 2). 'The fact that the MARC standard has been around for so many years has contributed to the creation of consistent and invaluable information within libraries. Such information is difficult to share and exchange with external entities. In an effort to make such data useful in the ever-evolving online environment, libraries across the world are currently experimenting with linked data technologies' (Papadakis, Kyprianos and Stefanidakis 2015, 2).

the form of the name that the archivist might choose for the local AAI. The AAI entry can then be associated with a chosen NAI-UID index entry (e.g. Thomas Hodgkin). Because the AAI is a reference to a document held at that archive, the AAI-UID can also (possibly) be attached to the Dublin Core (or equivalent) metadata record for the Record item,⁷⁷ enabling a link to be made from the AAI-UID entry to (for example) Google Scholar. But this is speculative and subject to further study and consideration by others. Every incidence of person names in the Record should be allocated a UID.⁷⁸ Names can (and often do) have variants in archival Records (Thomas Hodgkin, Tom Hodgkin, Thos Hodgkin, for instance). In research this problem is usually resolved by the construction of a local archival 'Authority Index' and this should remain the case. Many archives already have established name-based indexing systems and name-based indexes as a component of local search engines.⁷⁹

⁷⁷ 'Alternatively, they and librarians working with them might use cataloging tools based on international cataloging standards such as the Dublin Core Metadata Element Set, which comprises 15 key properties for describing resources. A variety of repository applications, including Omeka, support the Dublin Core. Because it conforms to the World Wide Web Consortium's (W3C) Resource Description Framework (RDF), catalogs of classical works that adhere to the Dublin Core standard can technically interoperate within the so-called semantic web of linked data. That is, when scholars, librarians, and others identify entities such as concepts or names of people and places in catalogs with uniform resource names (URNs), they can use standard web protocols to connect or "link" them together' (Waters 2023, 91).

⁷⁸ Digital representations of many, but not all, person name appearances in documents at an archive already exist in digitised full-body texts, indexes and catalogues and they can be found by using search engines via metadata and full-text searches (especially using AI). Many documents also have names and lists and tables of names embedded in the body of the text, and unless the whole of the text is digitised these names will not be machine readable, they will be hidden from digital view and therefore names must be collected manually. Also, from the nineteenth century onwards references began to commonly appear at the end of books and helpfully these provide tables of names appearing in texts, although these too are not necessarily discoverable digitally.

⁷⁹ 'While there is little consensus today about what to serve as information related to an authority entity, the mechanism for accessing the information via the HTTP protocol is currently well established. The technology involved is advanced and all kinds of library applications may benefit from the availability of library information on the web as linked data' (Papadakis, Kyprianos, and Stefanidakis 2015, 9).

2.7.3 The Research Authority Index

Lastly, the NAI-UID system must extend to include researcher datasets. Every researcher is free to (and must) create their own Research Authority Index when building datasets where the key field is person name, but RAIs must also become a part of the NAI-UID system. This will help reduce confusion when datasets in a repository with many different RAIs are later merged or linked. Including research datasets in the NAI-UID system both facilitates and harmonises the interoperability of researcher datasets (Figure 2.8).

Research Database Authority Index	Identifiers	Location
Quakers and activism	4567890123	Db a xxx repository
Thomas Hodgkin MD - RAI	3456789012	Db key field
Thomas Hodgkin NAI	1234567890	GRO - UK
Dr Hodgkin (Archive)	2345678901	TNA
Embedded in Document	Dublin Core UUID	TNA

Researcher (database) AI primary key record attaching primary key UUID to NAI UUID and Archive UUID (location)

Each research database may hold many records for person (n) and so each DB will have its own RAI. (e.g. individual DB records might be for T Hodgkin, Tom Hodgkin etc – each with their own UUID).

4

Figure 2.8 The research dataset component of the proposed National Authority Index system

First, the dataset itself is allocated a UID so that it can be found at the repository. The person name recorded in the dataset RAI (the key field) here is 'Thomas Hodgkin MD'. The researcher is studying Doctors of Medicine and prefers to use that form. The RAI entry is allocated a UID. The record referent UID at archive level (the AAI-UID) is also attached to the dataset record to indicate where the name was cited in the Record. A suggested NAI-UID allocation is also made.

It is important to note here that different archivists or researchers may disagree with the NAI-UID allocations made by others for what is thought to be the same person, and so the archivist or researcher is free to make an alternative allocation based on their own understanding of the Record information and the person's identity. This both embraces the unavoidable feature of 'messy data' in sources and allows archivists and researchers the freedom to argue their own interpretation of sources, through the allocations of AAls and RAIs.

2.7.4 The NAI-UID and the Semantic Web

Referencing the NAI-UID with its genealogical record information will help to place the record of the person's name within a family hierarchy, extending the scope of the NAI-UID system to embrace genealogy. By combining genealogical records with their family-based structure together with archival Records indexed via archival standards, it is likely that the increased combined information available to archivists and researchers will help reduce the number of misallocations. Machine reading algorithms can assist with data cleaning renditions of past person names across many datasets to speed up the transition to an NAI-UID system in the archives. The deployment of machine reading algorithms can take place alongside the deployment of trained volunteers (especially genealogists and researchers) in building out the NAI-UID system.

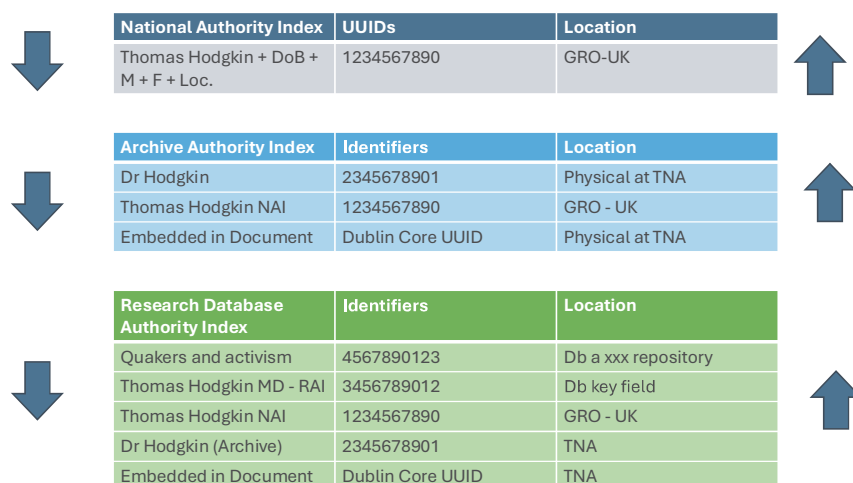


Figure 2.9 The complete NAI-UID system of indexes and UIs – the system works ‘top down’ and ‘bottom up’

The NAI-UID system links the NAI-UID index, AAI and RAI (Figure 2.9). The NAI-UID index UID is common to all three levels. Name variant preferences are facilitated at all levels. Pathways are facilitated from archival and research databases to the archive Record (and linking to Dublin Core is suggested). Horizontal pathways are also facilitated from one AAI or RAI to another and one EBPI family record to another (through genealogy). FAIR principles – ‘Findable, Accessible, Interoperable and Reusable’ – data affordances would be significantly enhanced through the disciplined use of common UIs and the NAI-UID system.

2.7.5 The NAI-UID system and the Case Studies

What if the NAI-UID system of UIs had been fully operable when the CEDA project discussed in the second part of this thesis was undertaken?

- Information gathering could have been quicker and systematised using the NAI-UID system.
- Data could have been collected from many archives and databases (for example The National Archives, the British Library, the Bodleian Library and the Wellcome Institute in the UK) using AI algorithms.
- Data found in other databases could have been more easily and confidently integrated into the project. Integrating datasets requires a high level of trust in the respective datasets and clear pathways from the integrated dataset to the respective sources on an individual record basis.
- The project database would have been easily findable and reusable by other researchers through the NAI-UID search engine.
- The 10,000 known name attributes revealed by the CEDA study could have been used by AI to identify even more enriching data, findable in other archives.

In summary, rich social network analysis and prosopography need not take years to research, and deeper structures and patterns of social interaction could have emerged over the forty-year timespan investigated by this study.

2.7.6 The integrated NAI-UID system

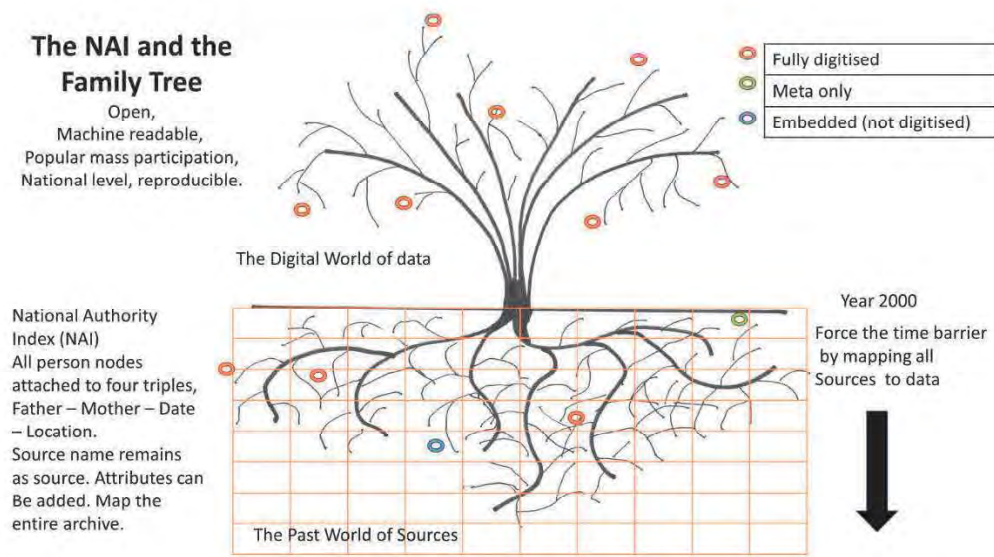


Figure 2.10 The NAI-UID system: the past world of ‘information’ becomes ‘data’ in the digital world

Figure 2.10 illustrates the entire NAI-UID concept in one image by linking the EBP world of the past with the Record world of the past. The diagram considers the world to be in two halves separated in time by digitisation, with the branches and leaves of the tree of knowledge appearing around the year 2000 into an emergent digital world. Researchers from today onwards might first consider using data in research projects that can be readily found in digital form and only sometimes seek information in non-digital forms (directly from Records). This, then, is a hybrid world with mixed digital and pre-digital information, and over time digital data will inevitably become more commonly used in research than pre-digital information in Records, both as digital data preponderates and as the techniques required in handling Records dwindle.

As we move deeper into the past (the years before widespread digitisation), prosopographical information becomes mixed between future digital (EBPD) and past

physical sources (EBPI), until before the last quarter of the twentieth century prosopographical information becomes solely embedded in physical Records. Little effort has been put into the digitisation of embedded person names and prosopographical information distributed throughout past Records. For instance, person name indexes frequently appear at the end of published documents (at least from the nineteenth century onwards), but these are not usually electronically searchable using archival search engines. This thesis argues that they should be. Moreover, archival search engines are not usually accessible at the level of the internet.

2.7.7 Triples, data tables, XML and the NAI-UID system

‘Triples’⁸⁰ in the form set out in the NAI-UID system are machine readable as Resource Description Framework (RDF) entities,⁸¹ allowing the NAI-UID system to make a major contribution to the Semantic Web (see Figure 2.11).⁸²

⁸⁰ <https://www.w3.org/TR/PR-rdf-syntax> (Accessed 27 May 2024).

⁸¹ ‘RDF properties can be defined with a domain, which indicates the subject of a triple is a member of a specified class. Similarly, defining a range of a property indicates that the object of a triple is a member of a specified class. Domains and ranges serve a dual purpose: first, to guide implementers as to how a particular property should be used, and second, to allow processing tools to derive new RDF relationships out of the connections implied by these definitions. For example, if the property *createdBy* is defined with a domain of *Book* and a range of *Person*, a system encountering this triple can assume the subject is of class *Book* and the object is of class *Person*, even if there are no known explicit RDF triples making those claims’ (Riley 2017, 11). See also (Allemang and Hendler 2011a, 49).

⁸² <https://www.w3.org/TR/WD-rdf-schema> (Accessed 27 May 2024).

A Simple Graph for Shakespeare's *The Tempest*

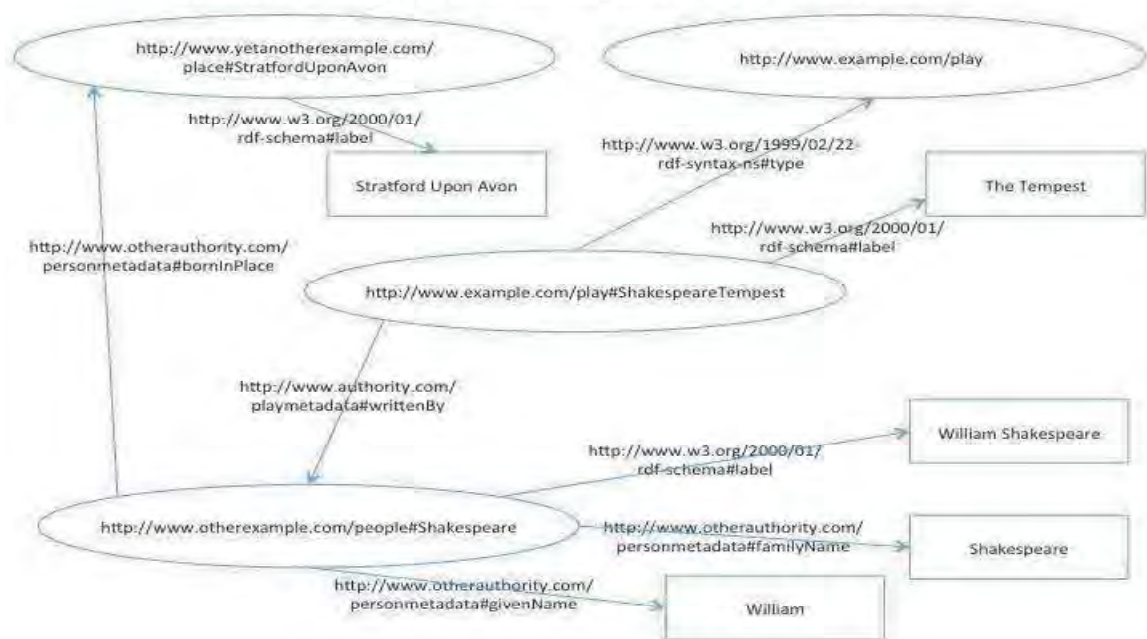


Figure 2.11 Example of a Resource Description Framework Schema (RDFS) (Riley 2017, 12)

Triples of EBPB (name, index integer) can be found and digitally compiled in all documents in the archive where EBPB is found.⁸³ Without an NAI-UID system or equivalent, these millions of person names will remain unfound. A system of authority indexes, metadata integration through Dublin Core and the wide use of UIs in the form of Digital Object

⁸³ 'In its simplest form, the concept could be applied to a sentence like "John drinks tea" or "David likes apples". Here, "John" and "David" are the subjects. The predicates are "drinks" and "likes" and the objects are "tea" and "apples". The idea of RDF triples builds on this: essentially RDF triples tie related resources and data together by indicating what something is, what attribute it has and how the attribute relates to it. A scientific paper can use RDF triples to express associated bibliographic information, whilst the relationships between individual tests, experiments, and their results can also be linked. Linking things in this way enables computers to pull up relevant data and results from all over the internet.

However, the power of RDF triples goes beyond linking specific words or phrases. Any of the parts of an RDF triple can be replaced with URIs, which are unique to a particular thing or concept. In the simple triple example "David likes apples", it would not be clear to a machine whether "apples" refers to the fruit or the computer, leading to ambiguity and irrelevant terms appearing in a semantic search. The distinction can be achieved by replacing the literal "apples" with a URI to either the fruit (dbpedia.org/page/Apple) or the computer company (dbpedia.org/page/Apple_Inc).

The use of URIs also allows established ontologies and vocabularies to be built. For instance, when describing a website, it is possible to make use of both Dublin Core metadata elements for describing resources and Library of Congress Subject Headings. Using established ontologies (or devising and making public your own ontologies if none exist already in your field) helps computers to be able to find all the related information.' <https://www.researchinformation.info/feature/rdf-triples-make-web-connections> (Accessed 27 May 2024).

Identifiers (DOIs) or Archival Resource Keys (ARKs) will make EBPD Record information easier to find in the form of triples compiled in authority indexes.⁸⁴ For example, the ARK Alliance is used by many libraries to uniquely identify documents and generate metadata (see Figure 2.12).

What are ARKs used for?

- genealogical records (8 billion [FamilySearch](#))
- publisher content (100 million [Portico](#))
- scientific datasets and records (22 million [INIST](#))
- scanned books and texts 30 million [Internet Archive](#))
- bibliographic records (15 million [BnF main catalog](#))
- museum specimens (15 million [Smithsonian Institution](#))
- public health documents (15 million [UCSF IDL](#))
- historical documents (21 million [CDL](#), 5 million [BnF Gallica](#))
- historical authors and scholars (4 million [SNAC](#))
- fine art museum collections (483,000 [Louvre](#))
- vocabulary terms (9,000 [Periodo](#), [YAMZ](#))

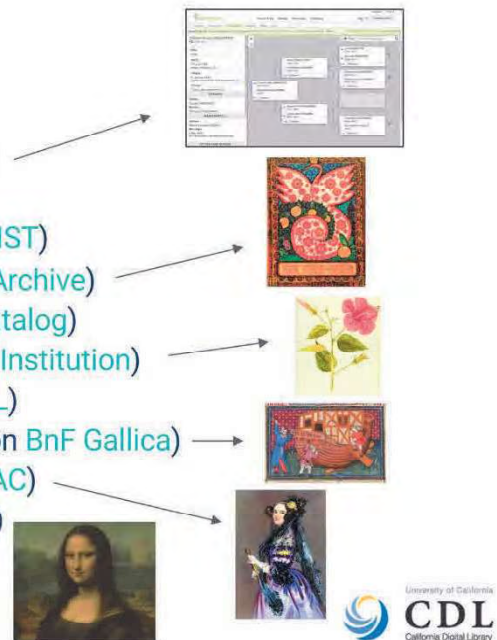


Figure 2.12 The ARK Alliance: 20 years, 850 institutions, 8.2 billion persistent identifiers

(<https://www.slideshare.net/jakkb1/the-ark-alliance-20-years-850-institutions-82-billion-persistent-identifiers-20211022>, Accessed 12 September 2023)

⁸⁴ 'Libraries keep their RDF data in lod-specific information systems commonly called triplestores. Triplestores are specialized database management systems for the storage and retrieval of RDF data (Rusher, 2010). Currently, many triplestores exist that serve different needs and demands (see Large Triple Stores). Some common triplestore engines are AllegroGraph, Virtuoso Universal Server, and Garlik 4store' (Papadakis, Kyprianos, and Stefanidakis 2015, 9).

The triple is also a core element of an RDF⁸⁵ and is likely to become a central organising feature of the Semantic Web (Version 2.0 or 3.0).⁸⁶ So working with triples in databases, as is common today, is likely to continue to be compatible with future digital data structures and infrastructures. Similarly, data tables commonly used in organising data in datasets of EBPD are also likely to be future compatible because conceptually the data table is at the heart of digital data structuring. So too is the use of hierarchical structures like XML which provide containers for metadata schemas⁸⁷ as well as acting as a data transit mechanism.⁸⁸ XML-type structures are a central feature of genealogical digitisation where genealogical data is built into family tree structures (see Figure 2.13).

⁸⁵ 'RDF is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed. RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a "triple"). Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications. This linking structure forms a directed, labeled graph, where the edges represent the named link between two resources, represented by the graph nodes. This graph view is the easiest possible mental model for RDF and is often used in easy-to-understand visual explanations.'

<https://www.w3.org/RDF/> (Accessed 26 September 2023).

⁸⁶ 'The World Wide Web is possible because a set of widely established standards guarantees interoperability at various levels. Until now, the Web has been designed for direct human processing, but the next-generation Web, which Tim Berners-Lee and others call the "Semantic Web," aims at machine-processible information. The Semantic Web will enable intelligent services—such as information brokers, search agents, and information filters—which offer greater functionality and interoperability than current stand-alone services. The Semantic Web will only be possible once further levels of interoperability have been established. Standards must be defined not only for the syntactic form of documents, but also for their semantic content. Notable among recent W3C standardization efforts are XML/XML schema and RDF/RDF schema, which facilitate semantic interoperability' (Decker et al. 2000, 63).

⁸⁷ 'Metadata Container Specifications: Since XML is human as well as machine readable, it is the preferred method for specifying metadata containers; it is self-descriptive. The container specifications, however, don't specify a single XML schema containing the complete set of metadata elements. Rather, they are frameworks of high-level elements that define extension points where specific descriptive, administrative, technical, and structural metadata can be embedded. This specific metadata is captured in extension schemas that define the specific metadata elements. It may be physically embedded or reference externally stored metadata' (Dappert and Enders 2010, 9-10).

⁸⁸ 'Many computer systems contain data in incompatible formats. Exchanging data between incompatible systems (or upgraded systems) is a time-consuming task for web developers. Large amounts of data must be converted, and incompatible data is often lost. XML stores data in plain text format. This provides a software- and hardware-independent way of storing, transporting, and sharing data. XML also makes it easier to expand or upgrade to new operating systems, new applications, or new browsers, without losing data. With XML, data can be available to all kinds of "reading machines" like people, computers, voice machines, news feeds, etc.'

https://www.w3schools.com/xml/xml_what.asp (Accessed 14 September 2023).

A Simple XML Record Example

```
<?xml version="1.0" encoding="UTF-8"?>
<work type="play">
  <workName>The Tempest</workName>
  <writtenBy>
    <playwright>
      <playwrightName>William Shakespeare</playwrightName>
      <bornInPlace>Stratford Upon Avon</bornInPlace>
    </playwright>
  </writtenBy>
</work>
```

<work>, <playwright>, and <bornInPlace> are examples of XML elements Stratford Upon Avon is an example of an element's value
type="play" is an example of an attribute name (type) and value (play)

Figure 2.13 A simple XML example (Riley 2017, 9)

The task of finding and preserving EBPd and the methodology proposed here to process and preserve it through the NAI-UID system recognise and exploit the deep structures of digital practice and digital organising concepts. The NAI-UID system is therefore a durable, reproducible, integrated and Semantic Web compatible system that can be made compliant with principles of good practice.⁸⁹

2.8 Chapter summary

To answer the question 'What is Evidence Based Prosopography and the National Authority Index system?', this chapter has offered an expanded explanation of EBPI and EBPd and their place within the concept of EBP and data science. The proposed NAI system has been fully described and illustrated with a working example taken from the P7 Case Studies. The

⁸⁹ <https://www.w3.org/Addressing/URL/uri-spec.html> (Accessed 4 February 2025).

discussion has considered issues in finding EBPI in archives and illustrated the academic value of EBP by outlining its use in the P7 Case Study exercises.

The chapter has taken a close look at some of the difficulties involved in working with EBP in the humanities. The devil is in the detail, because messy data, where provenance is difficult to establish, is common in DH, and messy data must be preserved in its messy state. Therefore, it would be wrong to attempt to restructure the evidence of the past to make it easier to manage digitally. The argument for the consideration of, as well as the possible adoption of, the NAI-UID infrastructure and its system has been set out as a core objective for DH.

The chapter has thus argued that EBP and the NAI-UID system are the way forward for digitisation in the humanities.

Chapter 3 EBP in the Digital Humanities

This chapter places Evidence Based Prosopography (EBP) in the context of rapidly developing Digital Humanities (DH). Information on Past Human Lives (PHL), central to the humanities, is explained, and how instead of focussing on information, DH has put greater emphasis on technology. Digitisation has advanced significantly in the sub-field of Digital History, but it will be shown that future progress here is limited unless the focus broadens to embrace information, data and its basic form, EBP. Developments in infrastructure and technologies are reviewed alongside the roles of researchers and their necessary technical supports, Research Software Engineers. Lastly the chapter answers the question:

Is infrastructure provision in the Digital Humanities sufficient to take up the national enterprise of the digitisation of Evidence Based Prosopography?

There are a great many partial and a few encompassing histories of Digital Humanities, and almost all of them are understandably grounded in the history of technology, given that the ‘digital’ part of the term is unambiguously technological.⁹⁰ This technological focus in digital history writing reinforces the general understanding that DH today is thought of as predominantly technological. This is problematic because it risks less regard being given to the role of other digital interpretations, in particular those based in data science, such as EBP.

⁹⁰ For a short general history of DH see (Hockey 2004); for a fuller and deeper historical analysis see (Castells 2011) which comprises a magisterial social sciences contemporaneous approach. Technologically rich histories are found in (Jones 2013), (Schreibman, Siemens, and Unsworth 2015), (Crymble 2021), (Kirschenbaum 2014) and (Thaller 2012). See also (Brennan 2018), (Breure, Doorn, and Boonstra 2006), (Ceci, Ferilli, and Poggi 2020), (de Groot 2020), (Friendly 2008a), (Gamble 2021), (Hilbert 2020), (Hering et al. 2014), (Liu 2013), (Marchese 2011), (I. Milligan 2022), (Piersma and Ribbens 2013), (Piotrowski and Fafinski 2020), (Scheinfeldt 2008), (Siebold and Valleriani 2022), (Spina 2021), (Sternfeld 2014), (Sula and Hill 2019), (Svensson 2016), (TANAKA 2022), (Thomas 2004) and (Zaagsma 2013).

Ludmilla Jordanova recognises that computers play many roles in modern academia,⁹¹ and the origins of the use of computers in DH has been traced back in time to Fr. Roberto Busa in 1949.⁹² However, Frank Owsley has also been identified as the father of DH in some recent histories, and he was also working in 1949. Crymble, writing in 2021, accepts both Busa and Owsley as foundational figures.⁹³ Either way, the origins of technology in DH go back around 75 years so computing in DH today is not a new phenomenon, although it is still considered by some as unpopular, as Luke Blaxill asserts in *‘Why do historians ignore digital analysis? Bring on the Luddites’* (Blaxill 2023). In 2017, in *The digital humanities and the digital modern*, James Smithies thought that DH was a marriage between ‘computing technology’ and history, if perhaps not such a happy one.⁹⁴ He made this appeal because he was concerned that DH was simply borrowing technologies from other fields without critically assessing the suitability of those technologies for humanities research (in particular

⁹¹ ‘There are many ways in which computers have come to play a part in the practice of history; they are a major component of history’s infrastructure. The Internet, by providing access to library catalogues and holdings worldwide, interactive websites and huge amounts of information, has become a particularly significant part of that infrastructure. The most robust educational systems provide sophisticated training in information technology’ (Jordanova 2019, 24).

⁹² ‘A common origin for the digital humanities and history is the remarkable insight [and salesmanship] of Roberto Busa, who, in 1949, convinced Thomas Watson to use the IBM Selective Sequence Electronic Calculator [SSEC] to find concordances in the writings of Saint Thomas Aquinas. The first demonstration occurred in 1952. This origin emphasises the computer as a tool that enhances a method of the humanistic research enterprise – that is, mining texts through pattern recognition and mapping’ (TANAKA 2022, 5).

⁹³ ‘Frank Owsley’s 1949 *Plain Folk of the South* ... Was it Owsley or Busa who led historians into computing? And did one influence “digital history” more than the other? Some scholars, who focus on the role of technology in the research process, see no difference between the two. Jane Winters (Britain) and Chad Gaffield (Canada) have independently suggested that both branches of research were part of a wider movement of progress that helped take historians from the state of mono- to interdisciplinary research’ (Crymble 2021, 18).

⁹⁴ ‘Researchers should consider searching for critical and methodological approaches to digital research in the humanities grounded in the nature of computing technology and capable of guiding technical development as well as critical and historical analysis’ (Smithies 2017, 2).

in history).⁹⁵ But in the same publication Smithies quotes Melissa Dinsman, for whom the origins of DH are harder to place.⁹⁶

What Smithies is referring to when he quotes Dinsman is commonly called the ‘Big Tent’ of DH. This concept became popular in 2011, when it was the theme of the DH2011 conference.⁹⁷ However, Melissa Terras was less convinced that the ‘Big Tent’ concept was helpful for DH because it risks diluting the disciplines involved through a lack of clear boundaries.⁹⁸

Chris Alen Sula and Heather V. Hill, in ‘The early history of digital humanities’, offer a cautionary reminder that in spite of the growth of the ‘Big Tent’ concept of DH, literature, languages and linguistics still dwarf all other disciplines, and history languishes far behind the top three (Table 3.1).

Discipline	All authors <i>N</i> (%) of authors	USA-affiliated authors <i>N</i> (%) of authors	USA doctorates, 1966–2004 <i>N</i> (%) of degrees
------------	--	---	--

⁹⁵ ‘[T]his will not suit all researchers, but it does benefit from a focus on a key tension wrought by the union of digital + humanities. At some stage it is necessary to accept that contemporary humanities research sometimes requires us not only to explain historical events and interpret texts, but to engineer working technical product to do so. It is a fantasy to suggest that we should rely solely on commercial or government software developers for our tools, or that humanists should reject research questions that require a computer. If we accept that, it follows that the research community should develop critical perspectives and methodologies that can support the interpretation and creation of software products’ (Smithies 2017, 4).

⁹⁶ ‘After multiple interviews with leading practitioners for an article in the *LA Review of Books*, Melissa Dinsman noted that DH are “large and increasingly indefinable even by those in its midst”, encompassing ‘computational research, digital reading and writing platforms, digital pedagogy, open-access publishing, augmented text, and literary databases ... media archaeology and theories of networks, gaming, and wares both hard and soft’ (Smithies 2017, 2).

⁹⁷ <http://dh2011.stanford.edu> (Accessed 3 June 2024).

⁹⁸ ‘The concept of a “big tent” to demarcate a group of individuals is a pragmatic and flexible description usually used to give strength in numbers, permitting a broad spectrum of views or approaches across the constituency. The term has been around for a long time, actually originating from religious American groups in the 19th century (see also “broad church”) rather than the circus background the name implies. It is most commonly applied to political coalitions that have a wide spread of backgrounds, approaches, and beliefs. In some respect it is well suited to Digital Humanities – what are the “Humanities” if not a “big tent” of scholars interested in the human condition and human society? What are the “Digital Humanities” if not a broad spectrum of academic approaches, loosely bound together with a shared interest in technology and humanistic research, in all its guises?’ (Terras 2016).

	CHum	LLC	CHum	LLC	
English and literature	136 (25.9)	104 (28.0)	94 (34.1)	41 (48.8)	51,092 (29.4)
Foreign languages	122 (23.2)	103 (27.7)	57 (20.7)	19 (22.6)	23,412 (13.5)
Linguistics	80 (15.2)	84 (22.6)	26 (9.4)	12 (14.4)	7,299 (4.2)
Arts and music	75 (14.3)	16 (4.3)	47 (17.0)	2 (2.4)	27,465 (15.8)
History	45 (8.6)	15 (4.0)	26 (9.4)	2 (2.4)	32,054 (18.4)
Religion and theology	4 (0.8)	8 (2.2)	1 (0.4)	–	16,031 (9.2)
Other humanities	57 (11.0)	42 (11.3)	25 (9.1)	8 (9.5)	16,568 (9.5)

Table 3.1 The dominance of literature, languages and linguistics in DH (Sula and Hill 2019, 199)⁹⁹

Further evidence of the ‘Big Tent’ concept in DH is the growth of digital research hubs which commonly coalesce around national or university-specific research centres (Breure, Doorn, and Boonstra 2006). In this regard, ‘Big Tent’ refers to a shared virtual space, since hubs in different locations network virtually and through networking express a common DH identity. The number of these centres has grown rapidly since 2006, and today the research centre hub has become a defining feature in DH.¹⁰⁰

Alongside the development of digital research hubs, most major international and regional libraries and archives have also actively developed digital strategies, commissioned digitisation projects and exploited the benefits of digitisation in the active management of collections; like research hubs, they also network virtually.¹⁰¹ For example, organisations

⁹⁹ ‘This article presents an empirical perspective on the early history of DH by tracing publications in two foundational journals (*Computers and the Humanities* (CHum), established in 1966, and *Literary and Linguistic Computing* (LLC)’ (Sula and Hill 2019, 199).

¹⁰⁰ ‘In the UK, mention should be made of the Arts and Humanities Data Service (aPHDs), which consists of five specialised centres for a variety of humanities fields distributed over the country. The Office for Humanities Communication at King’s College, London, is an umbrella organisation that fosters communication among scholars and others involved in computer-related projects and activities. It has published a series of monographs and collected papers concerned with the impact of computers in humanities scholarship and higher education. The Humanities Computing Unit at Oxford carries out research and develops resources in many areas of humanities computing. The unit provides support for academics in the humanities applying new technologies to their research and teaching. Among other things, the unit includes the Centre for Humanities Computing, the Oxford Text Archive, and the Humbul Humanities Hub. The mission of the Humanities Advanced Technology and Information Institute (HATII) at Glasgow University is to actively encourage the use of information technology and information to improve research and teaching in the arts and the humanities’ (Breure, Doorn, and Boonstra (2006, 14).

¹⁰¹ ‘The growing availability of digital historical sources is bringing about a change in the set-up and organisation of historical research. By bringing sources virtually together that are physically stored scattered in

such as the Arts and Humanities Research Council, the UK Data Service and the Collections Trust work with major libraries in the development of digital strategies in the UK.¹⁰²

3.1 Information in the Digital Humanities

A historiographical perspective might be more interested in offering a wider view, rather than just focusing on the adoption of digital technologies from the 1960s onwards as the sole defining feature of the digital in the humanities. There is abundant evidence of information gathering, information structuring and information analysis in wider society extending back over the last three millennia if the focus is put on information rather than technologies. In the seventeenth and eighteenth centuries home collections of books, antiquaria, flora and fauna, fossils, etc. flourished with the rise of the private library (Ciro 2002). Private collectors systematically sorted, categorised, catalogued and presented their collections in physical containers in what is now considered data table format.¹⁰³ As these first generations of modern collectors died, many bequeathed or placed their collections into family trusts or other forms of long-term preservation. National collections then began to arise to accommodate these bequests. In the UK, Hans Sloane's bequest of his private

archives around the globe (such as the VOC archives in Europe, Asia, and Africa), new opportunities for comparative research emerge, that were unfeasible in the past. Digital source collections can be studied from different perspectives by larger groups of researchers in the form of virtual "collaboratories" on the Web. Others can also check research if the digitised sources are made available' (Breure, Doorn, and Boonstra 2006, 19).

¹⁰² <https://collectionstrust.org.uk/digital-isnt-different/> (Accessed 24 March 2023).

¹⁰³ One of the principal features of ordered data, and it is almost universally so no matter where, when or what is being organised, is that the ordering will be a data table. Francis T. Marchese discusses how tabular representation of information developed between 13000 BCE and 1300 CE (Marchese 2011). From the seventeenth century country house libraries organised along their walls rows and columns of books in the form of a giant data table (Ciro 2002). Antiquarians also displayed their collections of artefacts in rows and columns in display cabinets. Ordered data is almost always presented in data table format (the graph is merely a pictorial rendering of a data table).

library resulted in the formation of the British Museum in 1753, although associated academic standards would not begin to emerge until after 1850, when the Public Libraries Act stimulated a popular interest in libraries (Morrish 2006).

In the nineteenth century, the use of statistical and ordered information became commonplace, not only in government¹⁰⁴ but also in the publications of reports and pamphlets of various societies and undertakings.¹⁰⁵ Much of this ordered information is in the form of embedded tables and lists and is available in the public domain archives.

Archival sources before the nineteenth century do contain embedded ordered information, although the further back in time the less common it is. Ordered information, especially when it contains evidence of PHL and where social networks can frequently be derived from it, is ideally suited to digital study and all sources at archives that include ordered information are of interest to EBP.

The concept of historical information organisation clearly precedes DH, and so it is reasonable to see DH as just a technological shift that perhaps gives too much emphasis to the new digital tools rather than the *longue durée* humanities interest in information. The practice of gathering, ordering and analysing information as data which digital humanities

¹⁰⁴ 'Two events in 1837 turned out to be the seeds of momentous changes to come. One was the installation of the first working telegraph in Britain. William Fothergill Cooke and Charles Wheatstone were contracted by Robert Stephenson, Engineer to the London and Birmingham Railway, to run a trial of their telegraph signalling system on Camden Bank, a sharp mile-long incline out of Euston station up which trains were pulled by mechanical cable. Their venture, together with those of other electrical pioneers, led on to the transatlantic telegraph cable, the telephone, radio, television, the Internet, and everything else that now carries the vast traffic of instant information that people and governments all over the world rely on to conduct their day-to-day business.

'The other key event of 1837 was the formation of the General Register Office for England and Wales, headed by a Registrar General appointed by the King. Though its primary role was administrative – to provide a reliable and legally effective system for registering births, marriages and deaths – the new Office was also a natural base for official statistics on the population: the Registrar General was required to prepare annual "abstracts" of the numbers of births, marriages and deaths, and was given the job of running the censuses of population' (Mahon 2009, 3).

¹⁰⁵ See the discussion in *Index, a history of the* about Jacques-Paul Migne's *Patrologia Latina* 1841–1845 (Duncan 2021, 206–208) and the Second Conference of Librarians in London in 1877 (Duncan 2021, 209).

encompasses today is arguably just the current phase in a global information management activity now more than 2000 years old.

3.2 The history of DH is characteristically technological

Many of the technologies used in DH have wider applications beyond the discipline (both in and out of academia) and also a longer history outside of DH than inside it. Hilbert gives an overview (Figure 3.1).

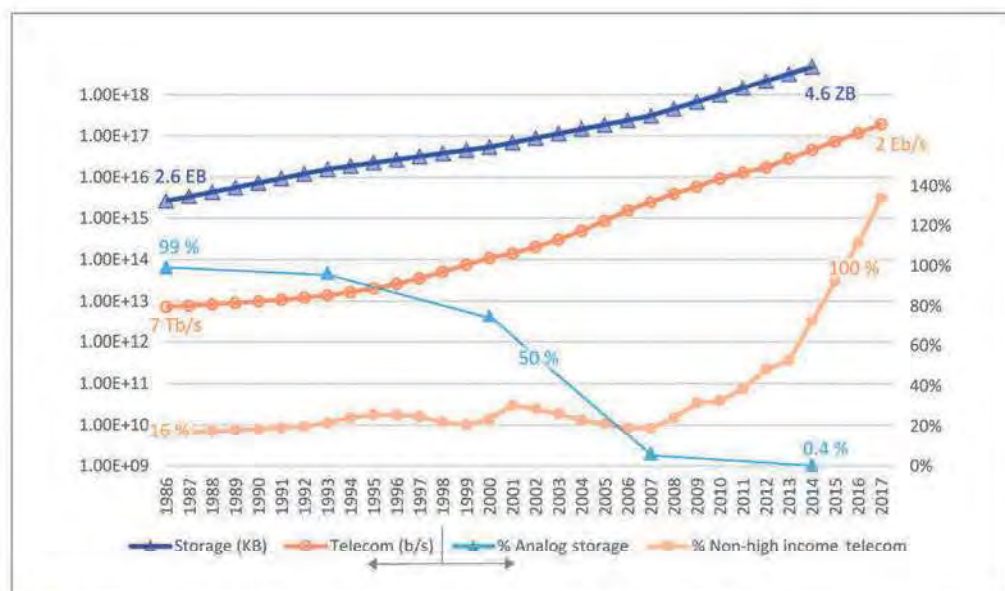


Figure 2. The world's technological capacity to store and telecommunicate information. Non-high-income telecom refers to the ratio of installed bandwidth capacity between non-high-income countries and high-income countries. EB, exabytes; ZB, zetabytes; Tb/s, terabits per second

Figure 3.1 Digital technology and social change: the digital transformation of society from a historical perspective (Hilbert 2020, 192)¹⁰⁶

¹⁰⁶ See also Hilbert (2020, 189): 'Digital technology, including its omnipresent connectedness and its powerful artificial intelligence, is the most recent long wave of humanity's socioeconomic evolution. The first technological revolutions go all the way back to the Stone, Bronze, and Iron Ages, when the transformation of material was the driving force in the Schumpeterian process of creative destruction. A second metaparadigm of societal modernization was dedicated to the transformation of energy (aka the "industrial revolutions"), including water, steam, electric, and combustion power. The current metaparadigm focuses on the

It is reasonable to say that DH practitioners have mostly borrowed and adopted a range of technologies from other, often much larger sectors (such as science, administration and business),¹⁰⁷ and it should be recognised that DH practitioners have benefited from their adoption. It is also important to recognise, however, that it is uncommon for DH researchers to use tools specifically designed for them. This places an added burden of creative thinking and adaption on DH, because DH researchers need to understand the disposition and theory of borrowed technologies in their native (non-DH) environments if they are to be critical of their applications as borrowed technologies in DH. It is therefore problematic to think of DH as having an independent technological history all of its own, and there is little evidence that early DH practitioners influenced technological tools development, with the singular exception of linguistics. Nevertheless, that an emerging discipline borrows from other disciplines in its early stages of development is not unusual,¹⁰⁸ and early pioneers of the practice of DH got to enjoy the freedoms that go with light disciplinary oversight as a trade-off for a lack of tool design responsibilities.

DH is always going to be a small player relative to many other digitised academies needing to find inventive and exploitative ways to succeed by taking and learning from the other sciences and also borrowing commercial technologies. Even so, the adoption of EBP, with its

transformation of information. Less than 1% of the world's technologically stored information was in digital format in the late 1980s, surpassing more than 99% by 2012. Every 2.5 to 3 years, humanity is able to store more information than since the beginning of civilization. The current age focuses on algorithms that automate the conversion of data into actionable knowledge.'

¹⁰⁷ Examples include databases, spreadsheets and graph databases.

¹⁰⁸ '[This chapter] clarifies how the rhetorical functions of borrowing from chaos theory trade on the newness of this field and the disciplinary prestige of the natural sciences. It concludes with some reflections on the problems that can arise when researchers borrow from disciplines as prestigious as the natural sciences. Throughout, the concern is with the persuasive functions served by borrowed knowledge' (Kellert 2008, Abstract).

wide interest in the digital study of EBPI and its representation as EBPD, has the potential to move DH from a minor and into a major role in academia.

This process of borrowing and adapting, rather than developing DH-specific technological capabilities in house and all of the difficulties that that entails, is perhaps behind the disappointing levels of take-up of the latter in DH. For instance, see Emmanuelle Delmas-Glass and Robert Sanderson for an account of the difficulties they see in finding supportive DH infrastructures, and the problems in digital take-up in academia which they hope future improvements in DH affordances will solve by integrating ‘communities of practice’.¹⁰⁹

In the case of Digital History, there has been a focus on tools borrowed from other disciplines often with a rush to use them without careful consideration of the nature and qualities of the object of study – the information as data in itself. This is perhaps why Digital History has struggled to live up to the expectations of many historians. This can be overcome by a change of focus onto the scientific study of PHL.

Chris Alen Sula and Heather V. Hill comprehensively illustrate the breadth of modern DH (in particular investigating one common characteristic widely held among DH practitioners, the idea that DH is a ‘Big Tent’), welcoming and embracing many academic schools (Figure 3.2).¹¹⁰

¹⁰⁹ ‘Researchers in all humanities disciplines have always developed and used analytic methods to approach their primary materials. Sometimes these are bespoke, developed by individual scholars for their own purposes. Sometimes they are shared between individuals, and a (relative) few are put to wider use within domains and communities. One characteristic they can be said to share, however, is that they are specific to particular kinds of research, or research questions. A method of parsing text in linguistics may have few obvious applications outside linguistics. The transformative potential of e-Science for these domains, however, is that it enables applications in different domains to be linked using common computational methods. In this paper, we have shown how these methods can be brought together as “methodological commons” and how scholarly primitives emerge across disciplines, and help build e-Infrastructures’ (Delmas-Glass and Sanderson 2020, 3794). See also Kitchin (2014).

¹¹⁰ ‘Most accounts of DH fail to chart an actual, historical course from humanities computing, with its alleged singular focus on the text, to present DH work in all its variety. Indeed, the existing histories balk when it comes to describing “big tent” DH work of the twenty-first century’ (Sula and Hill 2019, 191).

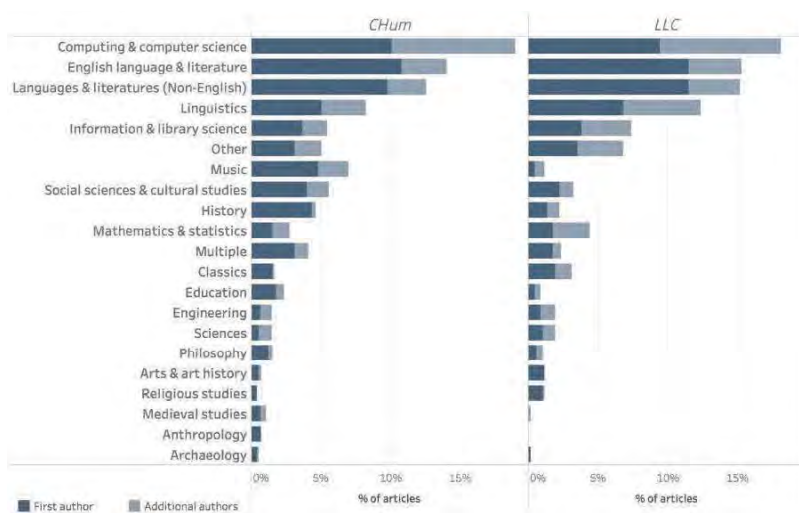


Figure 3.2 The early history of digital humanities: an analysis of computers and the humanities (1966–2004) and literary and linguistic computing (1986–2004) (Sula and Hill 2019, 193)¹¹¹

In 2000, John Unsworth introduced an important feature of DH which he called the ‘Methodological Commons’, introduced in ‘Scholarly primitives: what methods do humanities researchers have in common, and how might our tools reflect this’ (Unsworth 2000). Unsworth showed that DH research practices have features in common with those of all researchers: ‘discovering’, ‘annotating’, ‘comparing’, ‘referring’, ‘sampling’, ‘illustrating’ and ‘representing’. In the context of a study into EBP, it is difficult to see Unsworth’s insight here as more than a reductionist view, generally applicable to any and all human enquiries. However, at the time his characterisation will have been insightful to an academic area then unused to technological complexity in its practice. The Methodological Commons concept

¹¹¹ See also (Sula and Hill 2019, 193): ‘This study attempts to fill a gap in histories of DH by analysing publications in two key journals notable for their early presence in the field and their connections to prominent organizations. *CHum* was founded in 1966 as the official journal of the *Association for Computers and the Humanities* (ACH) until its final issue in 2004. *Literary and LLC* was founded by the *Association for Literary and Linguistic Computing* (ALLC) in 1986. From 2005 onward, it became the official journal of both the ALLC and the ACH. The journal was renamed *DSH: Digital Scholarship in the Humanities* in 2015 as an effort to rebrand to a wider audience (Vanhoutte, 2015b) and became the official journal of the *Association of Digital Humanities Organizations* (ADHO). The end date of the articles examined in this study reflects the final issue of *CHum* and predates wide circulation of *A Companion to Digital Humanities* (2004), which was important in shaping the field explicitly as DH (Terras, 2016).’

invited the idea of a Big Tent within it via the sheer inclusivity of the Methodological Commons (Figure 3.3).

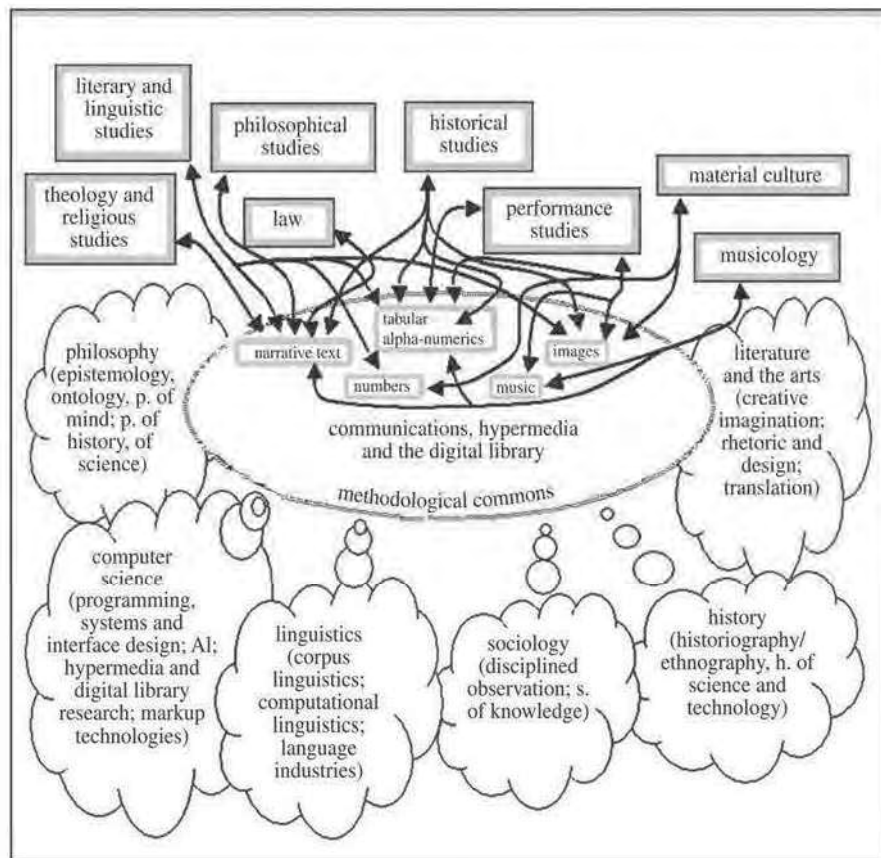


Figure 3.3 Digital humanities methodological commons (after McCarty & Short (2002), (Anderson, Blanke, and Dunn 2010, 3783).¹¹²

¹¹² 'Researchers within the digital humanities have over recent years sought to conceptualize and theorize their work so as to understand the complex relations between disciplinary practices, the digital materials which provide the sources for exploration, interpretation and analysis, and the methods and technologies that might be applied to answer research questions, and in the process to identify new research questions. McCarty & Short (2002) have developed an intellectual map to visualize these complex interactions around the concept of "methodological commons". Identifying and understanding the "commons" has proved a useful tool to explore methodological, epistemological, and normative divides between disciplines with a view to bridging those divides to better enable interdisciplinary work and to develop research infrastructures that support that work' (Anderson, Blanke, and Dunn 2010, 3782).

'McCarty & Short's (2002) map identifies three areas of mutuality. At the top are the disciplinary clusters denoting the range of research in the arts and humanities. Double-headed arrows to and from the methodological commons in the centre represent the connections to the types of content, tools, and methods most relevant to each. At the centre of the map is the methodological commons representing the different types of content most prevalent in the humanities (text, numeric and alpha-numeric, images, sound, etc.), the analytical tools and structures used to interpret and analyse that content, and the formal methods applied to interpret and analyse content. This representation is fluid, shaped by interdisciplinary engagement with the "clouds". Within the commons is also an unarticulated reference to the new forms of collaboration that this

A more important consequence of the adoption of the Methodological Commons concept is that it tends to divert focus away from another, equally important aspect of DH – the data itself.¹¹³ The ‘tools’-focused approach of the Methodological Commons used to explain the rising phenomenon of DH has since 2000 resulted in less thought being given to how non-technological aspects of DH studies such as a focus on data might influence digitisation in the future and what this might entail. This is especially the case as DH has quickly grown and extended its reach beyond linguistics and history and into wider academia.¹¹⁴ There has been much emphasis on understanding the nature, uncertainties and challenges of researchers and their tools, often at the expense of understanding the nature, uncertainties and challenges of acquiring and managing the information that researchers need.¹¹⁵ If, as

kind of interdisciplinarity requires. Below are “clouds” representing epistemological approaches and broad areas of disciplinary knowledge from both within and outside the arts and humanities with which scholars must engage in order to understand the arts and humanities e-research theoretically. McCarty & Short (2002) have used clouds to denote their nature as bodies of thought and the provisional understanding of their role in e-research. In this model mutual shaping takes place between the arts and humanities disciplines, the digital and non-digital source materials upon which they base their interpretations and analyses, the methods and tools applied to those materials, and the epistemological practices that are brought to bear to understand the interplay between the different elements of the commons. This is the space where interventions occur in the practices of knowledge creation that have the potential to provide new insights, and to lead to new e-research practices” (Anderson, Blanke and Dunn 2010, 3782–3783).

¹¹³ ‘So, one of the many things you can do with computers is something that I would call humanities computing, in which the computer is used as a tool for modelling humanities data and our understanding of it, and that activity is entirely distinct from using the computer when it models the typewriter, or the telephone, or the phonograph, or any of the many other things it can be’ (Unsworth 2002, 37).

¹¹⁴ ‘There are many ways of describing and understanding the digital humanities. I use the notion of “modes of engagement” as a means of describing the interrelation between the humanities and the digital. One important mode of engagement is technology as a tool, and much of the tradition of digital humanities has been built up around this mode: building archives, developing metadata schemes, creating, and using tools of different kinds, and focusing on methodology. Other modes of engagement include technology as an object of analysis and as an expressive medium. These modes of engagement are embedded in different epistemic traditions’ (Svensson 2016, 5).

¹¹⁵ ‘Digital history possesses a crucial set of common components—the capacity for play, manipulation, participation, and investigation by the reader. Dissemination in digital form makes the work of the scholar available for verification and examination; it also offers the reader the opportunity to experiment. He or she can test the interpretations of others, formulate new views, and mine the materials of the past for overlooked items and clues. The reader can immerse him/herself in the past, surrounded with the evidence, and make new associations. The goal of digital history might be to build environments that pull readers in less by the force of a linear argument than by the experience of total immersion and the curiosity to build connections.

well as focussing on researchers and their tools, attention were to be given to the object of research, the Records and their digital representations, then a more relevant and exciting picture of the future of DH might emerge.

If the universally understood characteristics of DH are that it is altogether a ‘Big Tent,’ ‘complex’ and ‘messy,’ then it is understandable that DH may require considerable infrastructural support for it to be of practical use to researchers. From a EBPD perspective, how, and to what extent, current DH infrastructures and disciplines might support EBP must come under critical scrutiny. This is because EBPD has a dual nature in that it manifests in two entangled forms: the digital representation and the related physical Record. EBPD must always be affixed to its Record to maintain the provenance of digital representations of EBPD overall.

Some writers have already turned their attention to a critical examination of the data itself.¹¹⁶ Anderson, Blanke and Dunn embrace the Methodological Commons concept, but go on to explain that the biggest challenge for DH is that ‘the humanities and arts face a “complexity deluge”’ in having to deal with ‘a multiplicity of types of information, much of it physically highly dispersed, sometimes difficult to find and complex to use’.¹¹⁷ EBP

(Versus the narrative anticipation of what comes next, this is a curiosity about what could be related to what and why.)’ (Cohen et al. 2008, 454).

¹¹⁶ ‘In both their promise and their threat, the digital humanities serve as a shadow play for a future form of the humanities that wishes to include what contemporary society values about the digital without losing its soul to other domains of knowledge work that have gone digital to stake their claim to that society. Or, precisely because the digital humanities are both functional and symbolic, a better metaphor would be something like the register in a computer’s central processor unit, where values stored in memory are loaded for rapid shuffling, manipulation, and testing—in this case, to try out new humanistic disciplinary identities evolved for today’s broader contention of knowledges and knowledge workers’ (Liu 2013, 410).

¹¹⁷ ‘By contrast, research work in the humanities is commonly characterized by the four Rs: reading, writing, reflection and rustication (Unsworth 2000, 2006) ... Rather than a data deluge, the humanities and arts face a “complexity deluge”, dealing with a multiplicity of types of information, much of it highly dispersed, difficult to find and complex to use; and instead of a focus on grids and HPC it needs tools and infrastructures that can take account of the essentially hermeneutic and practice-led nature of research practice in the humanities and arts’ (Anderson, Blanke, and Dunn 2010, 3781).

addresses and resolves that concern by systematising and ordering information critical to the study of PHL.

3.3 Digital History

Ever since E. H. Carr asked *What is history?* in 1961, the practice of history writing began to fragment and diversify away from the old Marxist/liberal duopoly, as a mood of pragmatism emerged across the discipline that eschewed grand theories and offered instead many disparate answers to the polemical question explicit in the title.¹¹⁸ Today, some fifty years later, Carr's question can be reutilised to ask 'What is Digital History?', especially now that the discipline has already decisively made the 'digital turn'.¹¹⁹ Douglas Seefeldt and William G. Thomas III provide a definition of Digital History that again focuses on its technological

¹¹⁸ 'So, when we write history (according to the Carr model) our motivation is disinterestedly to re-tell the events of the past with forms of explanation already in our minds created for us through our prior research in the archive. "Naturally" we are not slaves to one theory of social action or philosophy of history – unless we fall from objectivist grace to write history as an act of faith (presumably very few of us do this? Do you do this?). Instead we maintain our models are generally no more than "concepts" which aid our understanding of the evidence indeed, which grow out of the evidence. We insist our interpretations are independent of any self-serving theory or master narrative imposed or forced on the evidence. It is the "common sense" wish of the historian to establish the veracity and accuracy of the evidence, and then put it all into an interpretative fine focus by employing some organising concepts as we write it. We do it like this to discover the truth of the past' (Munslow 1997). See also (Tosh and Lang 2021, Preface xiii): 'history is a subject of practical social relevance; that the proper performance of its function depends on a receptive and discriminating attitude to other disciplines; and that the methods of academic history hold out the promise not of "truth" in any absolute sense, but of incremental growth in our knowledge of the past'. Jordanova's approach to historiography develops out of Carr's earlier philosophy: 'the pursuit of history is, whether practitioners acknowledge it or not, a political occupation ... This occurs in specific contexts that may be called "political" in that they involve decision making, the allocation of resources and esteem, and contested forms of public discourse ... it is extremely difficult to state briefly what the defining characteristics of history are, although an openness to many types of evidence and analytical flexibility are major features' (Jordanova 2019, 5–6).

¹¹⁹ 'The digital turn is thus transforming scholarship in three respects. The first is geography, as global projects – at least those drawing on repositories in the Global North, given the costs of digitization – are now possible in previously impossible ways. The second is digitization bias, the "Matthew Effect" of historical sources. The third is the transformation in the way in which sources are used, a shift from contextually aware skimming to surgical keyword search' (I. Milligan 2022, 28).

characteristics.¹²⁰ For Gerben Zaagsma, however, the concept of Digital History reflects only a transitional phase. At some point (soon) traditional narrative based history and Digital History will either merge as digital technology penetrates the world of historical sources (through digitisation or digital representation), or the skill set of historians widens to embrace both the worlds of sources and data.¹²¹

A historiographical analysis of the journey of history from 1960 to 2020 would be a major task, needing more time and space than this thesis can commit. Instead, the thesis considers only how DH is changing the study of history. ‘How (and what) are historians doing?’ by Charles Tilly (Tilly 1990) was written in the midpoint between Carr’s 1961 *What is history?* and today, and it provides a useful and better framework in which to place a review of recent and current practice in Digital History.

Tilly is writing in response to the ‘cultural turn’ and he focuses on social history.¹²² This is relevant to the work of this thesis because the basic unit of EBPD is the individual person, which is usually also the smallest unit of focus in social history. And like the P7 Case Studies

¹²⁰ ‘Digital history might be understood broadly as an approach to examining and representing the past that works with the new communication technologies of the computer, the internet network, and software systems. On one level, digital history is an open arena of scholarly production and communication, encompassing the development of new course materials and scholarly data collection efforts. On another level, digital history is a methodological approach framed by the hypertextual power of these technologies to make, define, query, and annotate associations in the human record of the past. To do digital history, then, is to digitize the past certainly, but it is much more than that. It is to create a framework through the technology for people to experience, read, and follow an argument about a major historical problem’ (Seefeldt and Thomas III 2009, 2).

¹²¹ ‘I would argue that there is no such thing as “digital history” as separate from “history” and I would hope that within a decade or so there will be no more talk of “digital history” as all history is somehow “digital” in terms of incorporation of new types of sources, methods and ways of dissemination (just as all humanities will be inherently “digital”). Nevertheless digital history is a transitional term that exists for a reason: it has helped to emphasise and put into focus new practices, whether in terms of analysis or knowledge (re)presentation or both; and it highlights how data and tools are changing historical knowledge production’ (Zaagsma 2013, 16).

¹²² Tilly writes just before the ‘cultural turn’, nevertheless his analysis holds good for the purposes of this thesis. See (Nash 2001).

in Chapter 6, for Tilly social history's main focus is also on the social group.¹²³ Tilly demarcates four modes of social history which create an axis with many persons at one end and the individual at the other. Tilly notes that individual contemporary historians might place their historical understanding somewhere along this scale rather than at the extremes, for example somewhere between (1) large social processes and (2) individual experiences.¹²⁴ He then combines these four modes with two characteristics representing the scale on which historians operate, to produce a graph distribution between 'large-scale horizons' and 'small-scale horizons' (see Figure 3.4).

¹²³ 'The division between social-scientific and other kinds of history reflects a much broader division within Western historical thinking. The division ultimately depends on philosophical choices which we might define provisionally as a series of alternatives:

1. History's dominant phenomena are (a) large social processes or (b) individual experiences.
2. Historical analysis centers on (a) systematic observation of human action or (b) interpretation of motives and meanings.
3. History and the social sciences are (a) the same enterprise or (b) quite distinct.
4. Historical writing should stress (a) explanation or (b) narrative' (Tilly 1990, 694).

¹²⁴ 'Rather than a strict dichotomy, to be sure, each of these pairs represents the poles of a continuum; the many historians who say "Let's look at the intersection between individual experiences and large social processes" or "Let's combine explanation with narrative" aim at the middle of those continua. Very few historians station themselves precisely at either pole of any continuum' (Tilly 1990, 694).

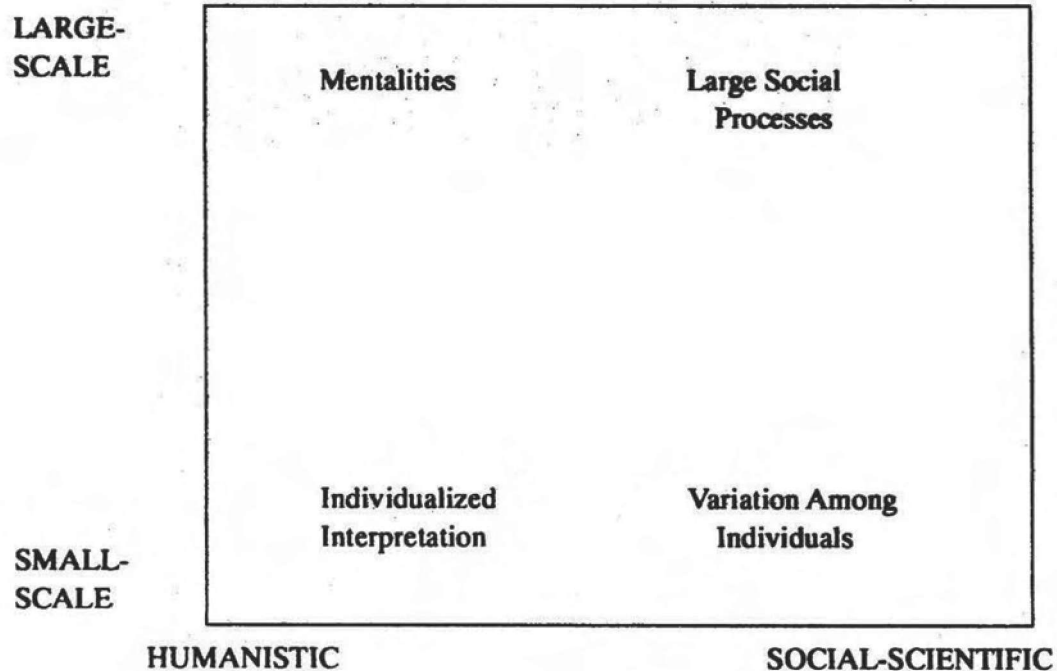


Figure 3.4 'A rough two-dimensional representation of variations in historical approaches looks like this' (Tilly 1990, 695)

Aware that this simplification is not without controversy among the community of historians because, he says, most historians would cluster their works in the centre of the chart, he nonetheless proceeds to place four 'exemplary historical works'¹²⁵ on the graph (see Figure 3.5).

¹²⁵ 'Let us explore the two-dimensional variation by reviewing some exemplary historical works—books that almost all historians will agree are excellent but that take very different approaches to their subjects. To see historical craftsmanship at work, let us concentrate on monographs rather than syntheses. To increase comparability and keep me on relatively certain ground, let us examine four outstanding works in western European and North American history: books by Carlo Ginzburg, E. P. Thompson, E. A. Wrigley and R. S. Schofield, and finally Olivier Zunz. The four do not constitute a representative sample of recent historical work—what four could? But they do provide relatively pure examples of monographs in each of the diagram's four corners, and thus mark out the space within which most historical work goes on' (Tilly 1990, 697).

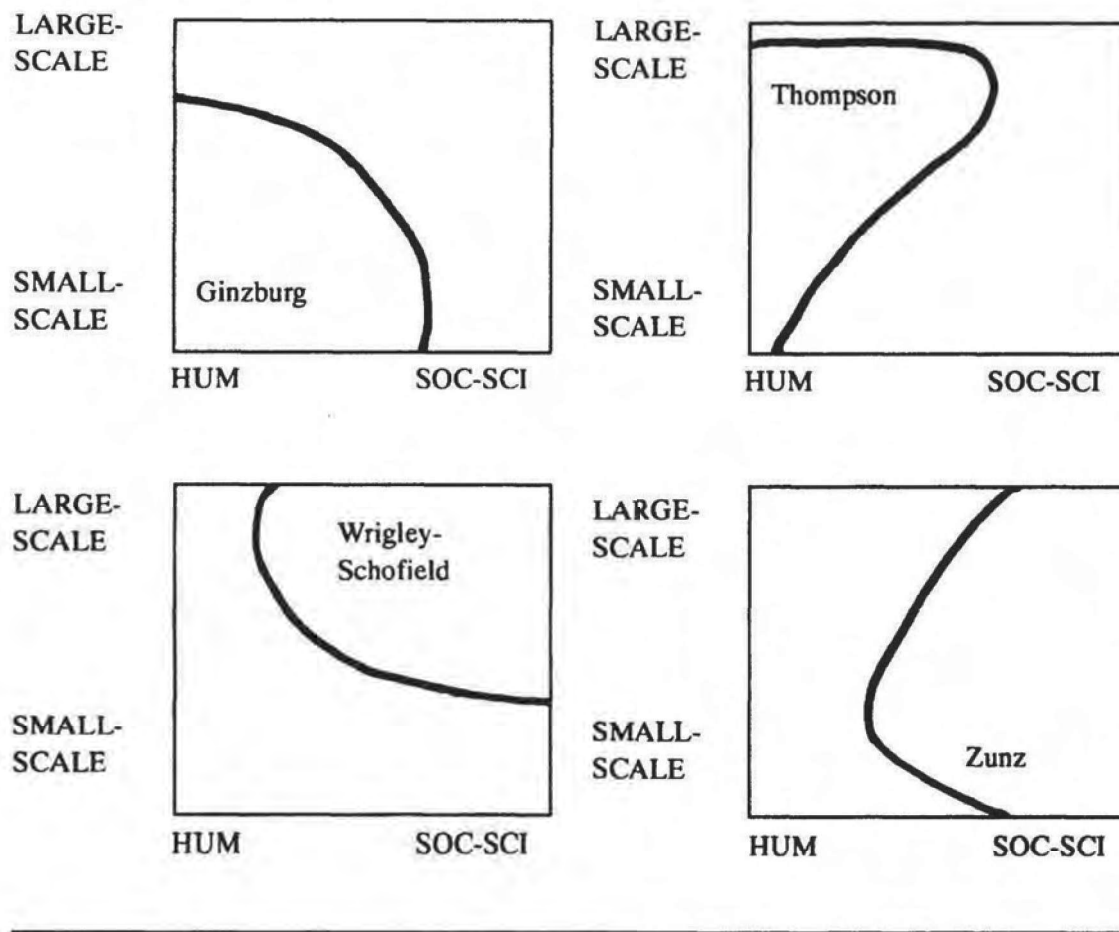


Figure 3.5 The monographs by Ginzburg, Thompson, Wrigley and Schofield, and Zunz

[They] fall far short of representing the full variety of Western historical work. Nevertheless, they provide relatively sharp examples of four distinctly different genres of historical research. Since none of the authors stays strictly in the corner assigned to him, we might represent the location of each this way. (Tilly 1990, 708)

Tilly places in the top right-hand corner of the graph E. A. Wrigley and R. S. Schofield's *The population history of England, 1541–1871: a reconstruction* from 1981. (Project Seven would appear in the bottom left.) Tilly outlines how different countries have exemplified his

model,¹²⁶ including French research groups and the Cambridge Group for the History of Population in the UK, of which Wrigley and Schofield were members.¹²⁷ He explains that large-scale sociological/scientific histories of this sort¹²⁸ were among the first to collect, use and rely on 'data' for the body of their work.¹²⁹ Finally, Tilly asks:

Is any synthesis of humanistic and social-scientific approach to history possible in principle? Yes, it is.

A resolution to the difficulty will arrive under one of four conditions:

- A discovery that reliable knowledge of human action is impossible, in which case both enterprises collapse.
- Proof that individual experiences are coherent and intelligible but large social processes are not, which condemns social science.
- Contrary proof that subjectivity is never reliably accessible but recurrent patterns of human action are, which scuttles humanistic history.

¹²⁶ 'Wrigley, Schofield, and their collaborators wrought their revolution by means of wide-ranging organization and a series of technical innovations. Their organization included the recruitment of volunteers throughout England who abstracted information about baptisms, burials, and marriages from more than 400 sets of local registers from as early as continuous series existed, and then shipped the information to Cambridge in standard format for computerization, tests for reliability, and aggregation into national estimates of annual numbers of births, deaths, and marriages. The central technical innovation was "back projection," the use of birth and death series to move back, 5 years at a time, from the sizes and age structures of populations enumerated in 19th-century censuses to best estimates of population sizes and age structures before that time' (Tilly 1990, 703).

¹²⁷ 'In the 1960s, both French and English demographers began to realize that the registers of baptisms, burials, and marriages long maintained by Christian churches would, under some conditions, yield reliable estimates of changes in the fertility, mortality, and nuptiality of the populations attached to those churches. In different ways, French and English research groups began the massive task of using those sources systematically to reconstruct vital trends before the age of regular national censuses, which began at the outset of the 19th century. The Cambridge Group for the History of Population and Social Structure took a threefold approach: extensive studies of household composition and other characteristics of local populations using whatever sources were available; derivation of refined estimates of vital rates by means of genealogies compiled from parish registers and similar records; and of national vital rates by aggregation of births, deaths, and marriages from a sample of parish estimates registers' (Tilly 1990, 703).

¹²⁸ 'Wrigley, Schofield, and the Cambridge Group carried out one of the largest enterprises, but not the only one. Philip Curtin's (1984) studies of the slave trade and of long-distance exchange in general, Robert Fogel and Stanley Engerman's (1974) econometric analyses of production under slavery in the United States, Jan de Vries' (1984) portrayal of European urbanization, Peter Lindert and Jeffery Williamson's (1983) analyses of changes in income and labor force during English industrialization, and Michael Schwartz's (1976) examination of smallholders' politics in the United States all exemplify the use of social scientific approaches to investigate history on the large scale' (Tilly 1990, 704-705).

¹²⁹ See (Tosh and Lang 2021, 55) for a suggestion of the beginnings of a data-led approach to historicism.

- Successful aggregation of reliably known individual experiences into collective action and durable social relations – which, if accomplished, will transform all the social sciences, as well as history. (Tilly 1990, 710)

The arguments made here by Tilly envision that in the future (he was speaking in 1990) there would be success in the ‘social history combined with humanism’ genre through the ‘aggregation of individual experiences’. Here Tilly prefigure this thesis’s interest in EBPD, because EBPD is concerned with the digital structuring of information necessary for the study of large numbers of PHL.

It can be anticipated from Tilly’s analysis that social history is now ready to embrace and explore EBPD. And it is through initiatives like NAI-UID that Zaagsma envisages the linking of information to representative data, allowing DH to develop new infrastructures and produce new information-based social histories. However, entrenched disciplines of practice in traditional schools of history are habitually slow to change and adapt and the weight of tradition can constrain innovation. The future acceptance and adoption of EBPD cannot be taken for granted.¹³⁰ This thesis recognises that the academic study of history is very broad, containing a large number of very different schools, and so EBPD will not be a pressing concern for all, as Hinke Piersma and Kees Ribbens make clear.¹³¹

¹³⁰ ‘Through the institutions that regulate academic disciplines in a variety of ways, conventions are formed about who counts as a historian and what counts as academic history ... in Britain there are two main national organisations: the Historical Association, founded in 1906, which embraces both professional and amateur historians, and the Royal Historical Society, founded in 1868’ (Jordanova 2019, 16-18).

¹³¹ ‘As a consequence of the aim of historians to basically cover all of human history, and knowledge being extracted from a wide array of sources, there is a large variety of methods that – depending on the research question and the nature of the source material – are used to gain insight into very different phenomena. Oral history is a very different way of gathering and interpreting sources than analysing medieval charters or interpreting ancient Greek potshards. Each source and each method of study inspires new questions, as is the case with digital historical research. But sources and methods also have their limitations’ (Piersma and Ribbens 2013, 83).

It should be noted that the 'Big Tent' characterisation of DH, along with its complexities and challenges, also arises at the level of individual disciplines as well as the higher level of the gathering of disciplines. Writing more recently than Tilly, Jordanova in 2019 recognises that history scholars' main product is still historical writing.¹³² She reminds us that while DH can be seen as a 'Big Tent', so also can the discipline of history itself. Many academic structures are today in a rapid process of change and realignment stimulated by digitisation. DH practitioners would do well to recognise that it is not just the emergence of DH that brings change but also the other factors recognised by Jordanova, each of which responds to pressures and calls for change deriving from both within and outside academia.¹³³

3.4 Prosopography

Prosopography¹³⁴ has a long history, perhaps beginning in 1898 with Theodor Mommsen's publication of the major collection of Latin inscriptions, *Corpus Inscriptionum Latinarum*

¹³² 'History may be thought of as a craft, requiring many skills, and at its heart sits writing' (Jordanova 2019, 13). See also (Champion 2017).

¹³³ 'If there is no activity or experience that is inherently outside historians' remit, then the discipline of history will necessarily find itself sharing concerns with other domains that study the human condition. Putting it this way suggests that the separations between disciplines derives not from any self-evident division in subject matter, but from other factors, such as custom and convention, interests and agendas, institutions and funding, values and beliefs' (Jordanova 2019, 67).

¹³⁴ 'There is no recorded history without humans and their stories; people as historical actors in events or occurrences in diverse contexts can be the subjects of written and oral histories. History is often concerned with what is regarded as the wealthy, the first or the famous as human subjects. Some historical actors meet these criteria, and accounts of their lives, careers, and roles in past events or occurrences have been prepared. Often, these actors were men and have been described as "great men" (Carnegie and Napier, 2012, p. 339; cf. Carnegie and Napier, 1996) or "heroes" (Hammond and Sikka, 1996, p. 79) or "movers and shakers" (Flesher and Flesher, 2003, p. 98), in accounting, business, and management and organisational history. The history of human subjects is biased towards the selected few, those notable or exceptional, but this does not constitute a complete history of people as contributors to human development and existence. Under-represented groups in history have been described as the "previously silent", indeed "voices that may have been deliberately silenced" (Napier and Carnegie, 1996, p. 5), for example "indigenous peoples" whose lands were colonised such as in Australia, Canada and New Zealand (Bastien et al., 2020; Carnegie, 2022; Mika et al., 2020), and "the lives of ordinary people" (Hammond and Sikka, 1996, p. 3) or as "the experiences of the 'voices from below'" (Carnegie and Napier, 2012, p. 346). In short, mainstream history is seen as insufficiently inclusive or too narrow'. (Carnegie and McBride 2023, 245)

(CIL). Mommsen initiated the accompanying *Prosopographia Imperii Romani* (PIR).¹³⁵ A hundred years later prosopography had moved from the compilation and publication of prosopographies in book form to the development and population of prosopographical databases with the *Prosopography of the Byzantine Empire* 1988. This move to the adoption of the database format placed prosopography amongst the early adopters of digital technologies and offered the possibility of widening, deepening and extending the scope of prosopographies, for example, by including in the dataset even greater numbers of individuals.¹³⁶ An important breakthrough was that unlike in hard copy publications where there was a need to fully reconcile ambiguities (especially in name authorities) because the hard copy format is rigidly inflexible, in the move to the digital prosopography could better embrace the ambiguities and uncertainties inherent in the data. A leading example of the changes that the new digital methodology brought is the *Prosopographie der mittelbyzantinischen Zeit* – PmbZ, where ordinary, non-elite persons could be included in the dataset.¹³⁷ The new methodology also opened up accessibility, both in terms of data and

¹³⁵ 'One of the first major endeavours to develop a formal prosopography was undertaken in response to the impact on the study of Roman history of the discovery of ever-increasing numbers of inscribed texts; a narrative which had been driven by the analysis of literary sources was suddenly confronted with a flood of data about individuals - some previously known, but many more newly revealed. It was the German scholar Theodor Mommsen (1817-1903), who was responsible for the organisation and publication of the major collection of Latin inscriptions, *Corpus Inscriptionum Latinarum* (CIL), who initiated the accompanying *Prosopographia Imperii Romani* (PIR). To begin with, this was partly an exercise in organising the data. The first edition of PIR, recording people in the Roman imperial period, appeared in 1898' (Roueché, Cameron, and Nelson 2023, 51)

¹³⁶ 'In 1988 it was noted that – 'there was no longer a rationale for excluding people. For Roman Egypt this meant tackling, for the first time, all the thousands of people briefly mentioned in the papyri. For the [Prosopography of the Byzantine Empire] PBE it meant abandoning the model which had continued from PIR to [Prosopography of the Later Roman Empire] PLRE - and which was imposed by print publication - of selecting only officeholders to record; this would come to transform the structure in which the information was presented.' (Roueché, Cameron, and Nelson 2023, 57)

¹³⁷ 'the huge crowd of people who could now be included in a digital resource were not most usefully presented in the traditional long article form; while this is entirely suitable for a collection of elite people, as in the early prosopographies, it is not a useful way to include butchers, bakers and candlestick makers. This kind of approach also meant processing the sources in a slightly different way. The researchers worked systematically through each source, recording all statements about individuals, and attaching such statements to person records. The information for an individual is not reconciled, as in an article-based prosopography,

users.¹³⁸ Microhistories developed in the area of business histories, where large numbers of people engaged in a single enterprise could be modelled and studied.¹³⁹ Europe is a major contributor to prosopography¹⁴⁰ but because this thesis is focussed on ‘evidence based prosopography’ attention here focusses on the development of the ‘factiod’ because ‘evidence based prosopography’ is a development of the ‘factiod’ model.

3.4.1 Prosopography at King’s College London and the ‘factiod’

Charlotte Roueché, Averil Cameron and Janet L. Nelson produced a comprehensive outline of the development of digital prosopography from the point of view of King’s College

and in the [*Prosopographie der mittelbyzantinischen Zeit*] PmbZ: though some sources are clearly more accurate than others, all available testimony is recorded. The PBE dataset is a guide to what is said in the sources; it has not set itself the task of source criticism, establishing which sources are more 'valuable', 'accurate' or 'true'. This was a radical new approach, made possible by working in a digital environment’ (Roueché, Cameron, and Nelson 2023, 59)

¹³⁸ ‘Digital prosopography emerged in the late 1990s and differs from previous prosopographical computing in being public facing, accessible online to a wide audience, and offering multiple ways co access data’ (Kowaleski 2021, 331)

¹³⁹ ‘Microhistory developed to enable historians to appreciate the lived experiences of people. The method was promoted by Italian scholars (Simone Cerutti, Carlo Ginzburg, Edoardo Grendi, Giovanni Levi, and Carlo Poni), who set up the journal *Quaderni storici* (Historical Notebooks) in 1966 and the series *microstorie* (Ginzburg, 1993). The approach soon spread with *Alltagsgeschichte* in Germany (Ludtke, 2003; Medick, 2001), post-Annales history in France (Lepetit, 1993; Revel, 1996; Rosental, 1996), and it was also adopted in North America (Demos, 1994; Ulrich, 1990)’. (Carnegie and McBride 2023, 251-252)

¹⁴⁰ ‘There have also been a number of later medieval digital projects emanating from Continental Europe in recent years, including prosopographies of students at the University of Paris (Projet Studium Parisiense), officers serving the Angevin royal dynasties (EuropAnge), papal legates in Hungary (Delegat Online), court personnel of the dukes of Burgundy (Prosopographia Curiae Burgundicae), and the entourage of Charles VI, king of France (Operation Charles VI)’. (Hammond 2021, 247) See also, ‘The Virtual Record Treasury of Ireland (VRTI)1 [1, 2, 3] is an all-island and international legacy from Ireland’s Decade of Centenaries’...‘in June 2022, the VRTI was launched’. ‘The VRTI-KG contains knowledge of notable People, Places, Offices, Organisations, and Interests and their interconnections, from the records of Irish history’. ‘In total, the VRTI-KG contains 8807 men and 965 women from Irish history uplifted from the DIB and Irish Exchequer Payments 1270-1446’. (Yaman et al. 2024, No page numbers). And see, BiographySampo - ‘The data was created by extracting knowledge from the underlying biographical texts, some 13100 short biographies published by the Finnish Literature Society, using natural language technologies. After this, the data was enriched by linking it to 13 external biographical databases, and to some additional collection databases of memory organizations and semantic web data services’. (Hyvönen et al. 2019, No Page numbers)

London,¹⁴¹ and in memory of John Bradley¹⁴² who was given the responsibility for undertaking and overseeing the transition [from INGRES to MySQL]. Prosopography at King's began in 1992 and led to the development of the digital *Prosopography of the Byzantine World*.¹⁴³ 'John [Bradley] was therefore involved over almost two decades in the development and delivery of several complex, similar but not identical, databases of people; among other important advances, this allowed him to develop and test the factoid model.' (Roueché, Cameron, and Nelson 2023, 63). Bradley led the development of the factoid model at King's and by 2010 the model had developed to embrace ambiguity and uncertainty in the data.¹⁴⁴ Matthew Hammond explains,

however, the document may be "lying" to us, or the person may have been claiming a tide which was not accepted universally, or the source may have been altered after the fact. That is why the "factoid" is only ever understood to be an assertion, a claim, as reflected in a given source. John Bradley and Michele Pasin have explained that the term "factoid" bears a deliberately ironic undertone for just this reason. Through the use of such factoids, the database could now represent assertions made in the sources without passing judgment as to their historical veracity. Factoids are at the heart of the prosopographical database because they constitute the moments where

¹⁴¹ 'The digital prosopographies at King's grew from a series of insights and interactions, which enabled the emergence of a profoundly new understanding of how to describe and record individuals in history. The creation of a Centre for such activities meant that a series of projects could be conceived not just within the boundaries of subject expertise, but as presenting a shared intellectual challenge'. (Roueché, Cameron, and Nelson 2023, 63)

¹⁴² 'I started work at King's College London (KCL) in 1997, first at its Centre for Computing in the Humanities which was subsequently renamed the Department of Digital Humanities (CCH/DDH)'. (J.D. Bradley 2020, 2)

¹⁴³ 'the founding of King's College London's Centre for Computing in the Humanities (CCH) in 1992, and the development there, with Dion Smythe, Harold Short, and, later, John Bradley, of what would eventually become the digital Prosopography of the Byzantine World project'. (Hammond 2021, 238)

¹⁴⁴ 'Bradley then further defined the factoid concept and its deployment: 'No factoids (including Events) appear unless they are linked both to Persons and to Sources. This principle is rigorously applied so that users are in a position to follow the Person-to-Source "trail", and to make their own reference to the relevant Source at any stage' (PASE 2010). (Roueché, Cameron, and Nelson 2023, 61)

entities connect. A factoid is situated at the nexus between Persons and Sources, and also often Places, but it may also connect with a potentially endless array of other sorts of information, such as sex, family, language, ethnicity, religion, education, titles, offices, landholdings, and events like marriages and deaths. The "factoid model" of digital prosopography has proven to be remarkably resilient and adaptable in reflecting the varying nature of historical sources. (Hammond 2021, 242-243)

In essence a factoid is an observable statement in a source about a specific individual. It is not a statement about a group of persons. A factoid is in large part defined by the requirements of the database, which is the intended location of the digital factoid. It must be attributed to (and therefore can be placed in the digital record of) an individual. It must also be an attribute of a person, capable of application to many – more than one – persons, and capable of standardisation, for example – occupation, age, sex, location. A factoid is declared a factoid by the observing researcher who needs to establish the determination of the suitability and applicability of the factoid. A factoid is interpreted and asserted as such and therefore is capable of challenge.^{145 146}

¹⁴⁵ 'A "factoid" refers to an assertion in a specific source about an individual. By calling the assertion a "factoid" and not a statement of "fact," this approach draws attention to the process required to extract information from a historical source in order to squeeze it into a structured format that computers can read. It highlights too the decisions or interpretative work that the editors of the database have to make about which assertions to include and how to structure them in the database'. (Kowaleski 2021, 319)

¹⁴⁶ 'Note that the aim of this and all factoid prosopographies is to make all assertions in sources available, even if they contradict one another. In other words, digital prosopographies are meant to be tools for research, not the final outcome of research on the people they record'. (319) (Kowaleski 2021, 319)

Recently King's have embraced Linked Open Data (LOD)¹⁴⁷ and the Semantic Web initiatives¹⁴⁸ with the 'Digital Prosopography of the Roman Republic (DPRR) project [which] has created a freely available structured prosopography of people from the Roman Republic'. (J.D. Bradley 2020, 1) This has produced the opportunity to link data to standard authority lists.¹⁴⁹

3.4.2 Issues in digital prosopography

Digital prosopographies are often difficult, time consuming enterprises, and costly. Two significant sustainability issues emerge. Firstly, due to the often long duration of time taken to develop the databases (indeed some very large databases have no project end terminus¹⁵⁰), maintaining and financing research effort over large timescales in an uncertain environment is problematic. Secondly, maintain and updating (either data or technology) to ensure that the considerable effort in the production phase of the prosopography is not later lost and that the prosopography can continue to be useful in the future, cannot always be guaranteed.¹⁵¹

¹⁴⁷ 'DPRR's data has also been made available as pure data, in a form suitable for LOD'...'we believed that DPRR connected in particularly useful ways to the three components of the idea of LOD: openness, linked, and data'. (J.D. Bradley 2020, 2)

¹⁴⁸ 'although almost all browser-mediated resources created at KCL have been open and freely available, they have not really been conceived as providing direct access to the data behind the web application'. (J.D. Bradley 2020, 2)

¹⁴⁹ 'how does DPRR's RDF server fit with one of the major interests from the Digital Humanities that have come out of LOD thinking: an interest in adding links from digital resources to standard authority lists such as V/AF[VIAF 2010-16]'. (J.D. Bradley 2020, 2)

¹⁵⁰ 'Our project does produce a profile view (figure 2), attached the URI for each individual, to that summarizes such characteristics as occupation, craft, civic office, and date range, and provides a convenient list of all records entered for that individual. But since our project is ongoing-there is no anticipated end since references [about] inhabitants of medieval London in online, print, and to manuscript sources number in the millions - the profile is not a static biographical dossier'. (Kowaleski 2021, 321)

¹⁵¹ 'A further problem, and one which affects all digital humanities resources to a certain extent, is the fact that the changing pace of technology means that it costs even more time, money, and expertise to keep such resources "on the road" in such a way that they will run on current versions of internet browsers and so forth. The fact that there is no research funding mechanism currently to provide for preservation and upkeep means

There is a natural tendency to use crowd sourcing approaches wherever both suitable and possible. This produces a temptation to compromise excellence in research with the needs to simplify to allow Independent researchers to make a considerable contribution.

Maryanne Kowaleski addressed these concerns in the development of *The Medieval Londoners Database* [(MLD)] using the Omeka S platform.¹⁵² (Kowaleski 2021)¹⁵³

A universal concern in digital prosopography is name authorities and (resulting from that) name linkages both with data within each prosopography and name linking one prosopography to others. The principle aim of this these is to address this concern and defect in the digital study of past human lives by the evidence based prosopography and the National Authority Index system. Maryanne Kowaleski explains,

‘Probably the biggest hurdles in prosopography, especially for medieval and other pre-modern populations, are the problems associated with name linkage, which in the medieval period are complicated by numerous variations in spelling for example, there was a John Clerk who was active in 1319 London and a John Clerc active in 1325, were they the same man? If they were both skimmers and both lived in Cheap ward, then the likelihood that they were the same person increases. In the MLD system, this linkage would be signalled by giving both of these John Clerks the same MLD Person_ID, which is a URI that can be read by any computer’...‘MLD allows both to be assigned the same Person_ID, but a "P" indicating that this is only a possible

that most, if not all, of these tools, which required vast amounts of money and person-hours to construct, will also disappear sooner or later’. (Hammond 2021, 248)

¹⁵² ‘The linked data capabilities of Omeka S also facilitate the use of fields such as Source_URI (Uniform Research Identifier) to connect to digitized versions of the source being used or Related_Link, which provides a direct URL to references to these individuals in other websites’. (Kowaleski 2021, 320)

¹⁵³ ‘keep data processing as simple as possible so that students and other researchers with little digital experience could participate in helping to structure data. Since bulk uploads of spread sheet data are easy in Omeka, we designed a basic data-entry spreadsheet that anyone can use, with columns mirroring the fields of the online database’. (Kowaleski 2021, 317)

link is placed in the Record_Type field, and the entire record is shaded pink in the online database remind to users that the link is tentative'. (Kowaleski 2021, 323-325)

In the absence of an independent and national name authority system Prosopographers must resort to shared, if informal, practise in name disambiguation.¹⁵⁴

3.5 Research infrastructure provision

What the term 'infrastructure' signifies in DH is still fluid.¹⁵⁵ Its origin probably reaches back to Susan Star and Karen Ruhleder in 'Steps toward an ecology of infrastructure: design and access for large information spaces'. They explain that infrastructures need not be fixed provisions, they can be temporary, tailored to the specific needs of each project.¹⁵⁶

¹⁵⁴ 'the monumental 83 volume nineteenth century Real-Encyclopaedie der classischen Altertumswissenschaft [Pauly et al 1893-] -referred to as RE and once called by a DPRR project member the "grandfather" of all DPRR's prosopographical sources. RE continues to provide the basis against which historical identity of individuals is argued even today'. (J.D. Bradley 2020, 3)

¹⁵⁵ 'By "digital infrastructure," I follow the Atkins report and Our Cultural Commonwealth and mean in this essay to denote the collection of standards, software, digital content, and expertise that directly supports scholarly research. Because "infrastructure" is a relative term, the scholarly infrastructure discussed here in turn depends on deeper layers of support. At one level, there are platforms of various kinds for digital search and messaging; other levels include networking and storage protocols and technology' (Waters 2023, 88).

¹⁵⁶ 'The tool emerges in situ. By analogy, infrastructure is something that emerges for people in practice, connected to activities and structures ... an infrastructure occurs when local practices are afforded by a larger-scale technology, which can then be used in a natural, ready-to-hand fashion. It becomes transparent as local variations are folded into organizational changes, and becomes an unambiguous home – for somebody. This is not a physical location nor a permanent one, but a working relation – since no home is universal. Experience with groupware suggests that highly structured applications for collaboration will fail to become integrated into local work practices (Ruhleder and Jordan, in progress). Rather, experimentation over time results in the emergence of a complex constellation of locally-tailored applications and repositories, combined with pockets of local knowledge and expertise. They begin to interweave themselves with elements of the formal infrastructure to create a unique and evolving hybrid. This evolution is facilitated by those elements of the formal structure which support the redefinition of local roles and the emergence of communities of practice around the intersection of specific technologies and types of problems. These observations suggest streams of research that continue to explore how infrastructures evolve over time, and how "formal," planned structure meld with or give way to "informal," locally-emergent structure' (Star and Ruhleder 1996, 5-6).

Star and Ruhleder were writing at a time when, due to the relative infancy of digital technologies, interfacing with technology was not easy and resistances could easily arise between individual researchers and the computer. Many of these human resistances still exist today, even though many of the interface barriers have, over time, been removed. Star and Ruhleder identify eight important key attributes of digital infrastructure which remain relevant:¹⁵⁷

- Embeddedness
- Transparency
- Reach or scope
- Learned as part of membership
- Links with conventions of practice
- Embodiment of standards

¹⁵⁷ '[I]nfrastructure emerges with the following dimensions:

- *Embeddedness*. Infrastructure is "sunk" into, inside of, other structures, social arrangements and technologies;
- *Transparency*. Infrastructure is transparent to use, in the sense that it does not have to be reinvented each time or assembled for each task, but invisibly supports those tasks;
- *Reach or scope*. This may be either spatial or temporal – infrastructure has reach beyond a single event or one-site practice;
- *Learned as part of membership*. The taken-for-grantedness of artifacts and organizational arrangements is a *sine qua non* of membership in a community of practice (Lave and Wenger, 1992; Star, in press). Strangers and outsiders encounter infrastructure as a target object to be learned about. New participants acquire a naturalized familiarity with its objects as they become members;
- *Links with conventions of practice*. Infrastructure both shapes and is shaped by the conventions of a community of practice, e.g. the ways that cycles of day night work are affected by and affect electrical power rates and needs. Generations of typists have learned the QWERTY keyboard; its limitations are inherited by the computer keyboard and thence by the design of today's computer furniture (Becker, 1982);
- *Embodiment of standards*. Modified by scope and often by conflicting conventions, infrastructure takes on transparency by plugging into other infrastructures and tools in a standardized fashion.
- *Built on an installed base*. Infrastructure does not grow *de novo*; it wrestles with the "inertia of the installed base" and inherits strengths and limitations from that base. Optical fibers run along old railroad lines; new systems are designed for backward-compatibility; and failing to account for these constraints may be fatal or distorting to new development processes (Monteiro, et al., 1994).
- *Becomes visible upon breakdown*. The normally invisible quality of working infrastructure becomes visible when it breaks: the server is down, the bridge washes out, there is a power blackout. Even when there are back-up mechanisms or procedures, their existence further highlights the now-visible infrastructure' (Star and Ruhleder 1996, 5).

- Built on an installed base
- Becomes visible upon breakdown

Tensions between drives for national standardisation and the desire to encourage innovation and adaptation at the local level were clearly identified and emphasised by Star and Ruhleder. There was and still is a tension which defines much of the current debate in DH concerning the roles and purposes of infrastructures. This tension is perhaps not capable of resolution, suggesting both that a balance needs to be struck between conforming to larger group discipline and individual and independent expression in research, and that this balance needs to be struck at the level and point of each individual research project, as it arises. This tension is revealed in today's differing attitudes to the drives for common infrastructure in DH, with scholars in the US leaning more towards facilitating individual research projects and scholars in the EU leaning towards EU-level standardisation and objective-focused co-operation. The UK is yet to find its way now that it is no longer a member of the EU: whether it will work alongside EU structures, adopt US independent practices or both is as yet unclear. The EBP system described here meets both structures, with the NAI authority index organised centrally allowing archives, genealogy affordances and individual research projects to use local authority files linked to the central NAI index.

3.5.1 International and national infrastructures

Organisations supporting and providing infrastructure exist at regional and national levels. At the regional level and based in the US, three organisations have a somewhat patchy

global reach: centerNet,¹⁵⁸ the Transatlantic Platform (T-Ap) which also reaches out from the US to Europe, and the Alliance of DH Organizations (ADHO).^{159,160}

In the US the National Endowment for the Humanities Office of Digital Humanities (NEH)¹⁶¹ provides umbrella support on a national basis, and a similar role is performed at the European level by the European Association for Digital Humanities (EADH),¹⁶² which sits underneath the European Research Area (ERA) umbrella.¹⁶³ There are also several EU

¹⁵⁸ '[C]enterNet is an international network of digital humanities centers formed for cooperative and collaborative action to benefit digital humanities and allied fields in general, and centers as humanities cyberinfrastructure in particular. Anchored by its new publication *DHCommons*, centerNet enables individual DH Centers to network internationally — sharing and building on projects, tools, staff, and expertise. Through initiatives such as Day(s) of DH and Resources for Starting and Sustaining DH Centers, centerNet provides a virtual DH center for isolated DH projects and platform for educating the broader scholarly community about Digital Humanities.' <https://www.transatlanticplatform.com> (Accessed 9 October 2023).

¹⁵⁹ 'The Alliance of Digital Humanities Organizations (ADHO) is an umbrella organization whose goals are to promote and support digital research and teaching across arts and humanities disciplines, drawing together humanists engaged in digital and computer-assisted research, teaching, creation, dissemination, and beyond, in all areas reflected by its diverse membership. ADHO supports initiatives for publication, presentation, collaboration, and training; recognizes and supports excellence in these endeavors; and acts as a community-based consultative and advisory force.' <https://adho.org> (Accessed 9 October 2023).

¹⁶⁰ 'We are resource specialists working in libraries and archival centres, humanities computing groups, and other professional arenas. We are academic administrators, and members of the private and public sectors. We are independent scholars, students, graduate students, and research assistants. We are from countries in every hemisphere.' The following organizations are current members of ADHO: Association for Computers and the Humanities (ACH); Australasian Association for Digital Humanities (aaDH); Canadian Society for Digital Humanities/Société canadienne des humanités numériques (CSDH/SCHN); centerNet; Digital Humanities Alliance for Research and Teaching Innovations (DHARTI); Digital Humanities Association of Southern Africa (DHASA); Association for Digital Humanities in the German Speaking Areas (DPHD); European Association for Digital Humanities (EADH); L'association francophone des humanités numériques/digitales (Humanistica); Japanese Association for Digital Humanities (JADH); Korean Association for Digital Humanities/한국디지털인문학협의회 (KADH); Red de Humanidades Digitales (RedPHD); Taiwanese Association for Digital Humanities (TADH). <https://adho.org/about/> (Accessed 07 October 2023).

¹⁶¹ <https://www.neh.gov/divisions/odh> (Accessed 9 October 2023).

¹⁶² 'The EADH brings together and represents the Digital Humanities in Europe across the entire spectrum of disciplines that research, develop, and apply digital humanities methods and technology. The EADH also supports the formation of DH interest groups in Europe that are defined by region, language, methodological focus or other criteria. The European Association for Digital Humanities (EADH) was founded in 1973 under the name Association for Literary and Linguistic Computing (ALLC) with the original purpose of supporting the application of computing in the study of language and literature. As the range of available and relevant computing techniques in the humanities increased, the interests of the association's members have broadened substantially and encompass not only text analysis and language corpora, but also history, art history, music, manuscript studies, image processing and electronic editions. The association's new name, which was adopted in 2012, reflects this significant widening of scope. Today the EADH's mission is to represent European Digital Humanities across all disciplines.' <https://eadh.org/about> (Accessed 18 October 2023).

¹⁶³ 'The European Research Area (ERA) is the ambition to create a single, borderless market for research, innovation and technology across the EU. It helps countries be more effective together, by strongly aligning

funding infrastructures such as Horizon Europe,¹⁶⁴ and yet other infrastructure providers focus on tool development and training in expertise: the Digital Research Infrastructure for Arts and Humanities (DARIAH)¹⁶⁵ and the Common Language Resources and Technology Infrastructure (CLARIN).¹⁶⁶ For a critique of ‘globalisation’ in DH infrastructure provision, and especially concerning inequality of provision, see ‘Infrastructuring digital humanities: on relational infrastructure and global reconfiguration of the field’ (Pawlicka-Deger 2021).

At the national level in the UK, infrastructure support is performed by the Joint Information Systems Committee (JISC), established on 1 April 1993 (see Figure 3.6).

their research policies and programmes. The free circulation of researchers and knowledge enables better cross-border cooperation, building of critical mass, continent-wide competition. ERA was launched in 2000 and a process to revitalise it began in 2018.’ https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/european-research-area_en (Accessed 7 October 2023).

¹⁶⁴ https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe_en (Accessed 7 October 2023).

¹⁶⁵ ‘The Digital Research Infrastructure for the Arts and Humanities (DARIAH) aims to enhance and support digitally-enabled research and teaching across the arts and humanities. DARIAH is a network of people, expertise, information, knowledge, content, methods, tools and technologies from its member countries. It develops, maintains and operates an infrastructure in support of ICT-based research practices and sustains researchers in using them to build, analyse and interpret digital resources. By working with communities of practice, DARIAH brings together individual state-of-the-art digital arts and humanities activities and scales their results to a European level. It preserves, provides access to and disseminates research that stems from these collaborations and ensures that best practices, methodological and technical standards are followed.’ <https://www.dariah.eu> (Accessed 7 October 2023). See also (Kaltenbrunner 2017).

¹⁶⁶ ‘CLARIN is a digital infrastructure which provides easy and sustainable access to a broad range of language data and tools to support research in the humanities and social sciences, and beyond. CLARIN provides access to multimodal digital language data (text, audio, video) and advanced tools with which to explore, analyse or combine these datasets.’ <https://www.clarin.eu> (Accessed 7 October 2023).

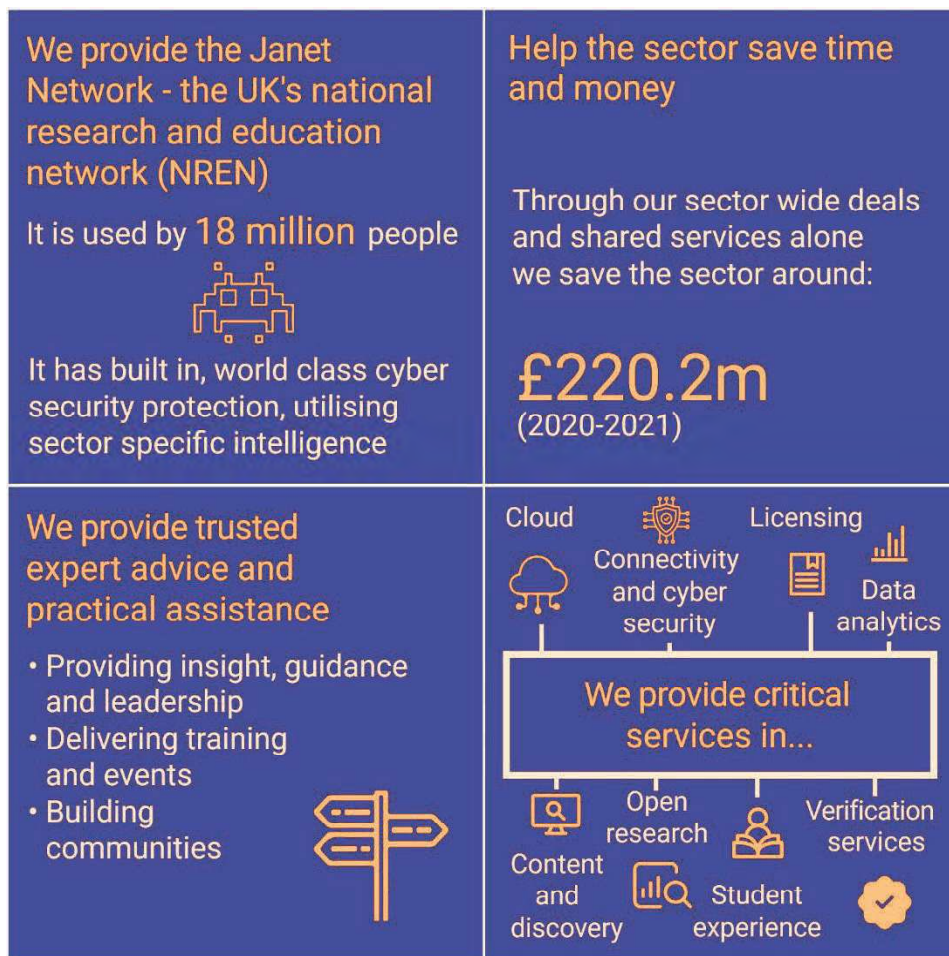


Figure 3.6 JISC scope and services (<https://www.jisc.ac.uk/about-us>, Accessed 1 June 2023)¹⁶⁷

Other infrastructure providers in the UK include The National Archives,¹⁶⁸ the British Library Web Archive,¹⁶⁹ the UK-Ireland Digital Humanities Association,¹⁷⁰ the UK Economic and

¹⁶⁷ See also <https://www.linkedin.com/company/jisc/?originalSubdomain=uk> (Accessed 1 June 2023): 'Jisc is the UK digital, data and technology agency focused on tertiary education, research and innovation. We are a not-for-profit organisation and believe education and research improves lives and that technology improves education and research. We provide managed and brokered products and services, enhanced with expertise and intelligence to provide sector leadership and enable digital transformation.'

¹⁶⁸ <https://www.nationalarchives.gov.uk> (Accessed 9 October 2023).

¹⁶⁹ <https://www.bl.uk/collection-guides/uk-web-archive> (Accessed 9 October 2023).

¹⁷⁰ 'The UK-Ireland Digital Humanities Association brings together researchers, practitioners and organisations from both countries to build a collaborative vision for the field, with a focus on issues such as sustainability, inclusivity, training, advocacy and career progression. This vision for the field builds on long-standing partnerships, research, and centres of excellence to further development and innovation in digital humanities. The Association seeks to nurture the capacity for excellent research and teaching in digital humanities, to establish and sustain more effective connections across sectors, and to create new pathways for collaboration.'

Social Research Council (ESRC),¹⁷¹ the Arts and Humanities Research Council (AHRC), a part of UK Research and Innovation (UKRI),¹⁷² and lastly the UK Data Service,¹⁷³ which funds research in the arts and humanities. Other smaller focused organisations also offer infrastructure and support, for example the Institute of Historical Research Digital Seminars.¹⁷⁴ Infrastructure support provided by these organisations comprises funding, networking, nurturing, sharing through conferences and interest groups, and some (especially the EU organisations) standardisation initiatives in ontology, taxonomy and methodologies. Publishing and journal infrastructure is also well provisioned (for example, the *International Journal of Humanities and Arts Computing* (IJHAC)¹⁷⁵ and *Digital Scholarship in the Humanities* (DSH)).¹⁷⁶

3.5.2 Google Scholar

These collaborations include partners in higher education; galleries, libraries, archives and museums; the technology sector; and the creative industries. Institutions and individuals engaged with the Association will work to create new and sustainable partnerships across Ireland and the UK, as well as with the international community.' <https://digitalhumanities-uk-ie.org> (Accessed 9 October 2023).

¹⁷¹ 'ESRC is the UK's largest funder of economic, social, behavioural and Past Human Data science.' <https://www.ukri.org/councils/esrc> (Accessed 7 January 2023).

¹⁷² 'We invest in research and innovation to enrich lives, drive economic growth, and create jobs and high-quality public services across the UK. We are transforming tomorrow together'. <https://www.ukri.org> (Accessed 7 October 2023).

¹⁷³ <https://ukdataservice.ac.uk/learning-hub/research-data-management> (Accessed 9 October 2023).

¹⁷⁴ https://www.youtube.com/channel/UCLBI7fD7EQmu652Pr_oWEYw (Accessed 10 October 2023).

¹⁷⁵ 'IJHAC: A Journal of Digital Humanities (formerly *History and Computing*) is one of the world's premier multi-disciplinary, peer-reviewed forums for research on all aspects of arts and humanities computing. It focuses both on conceptual or theoretical approaches and case studies or essays demonstrating how advanced information technologies further scholarly understanding of traditional topics in the arts and humanities. The journal also welcomes submissions on policy, epistemological, and pedagogical issues insofar as they relate directly to computing-based arts and humanities research.' <https://www.eupublishing.com/loi/ijhac> (Accessed 7 October 2023).

¹⁷⁶ 'DSH or *Digital Scholarship in the Humanities* is an international, peer reviewed journal which publishes original contributions on all aspects of digital scholarship in the Humanities including, but not limited to, the field of what is currently called the Digital Humanities.' <https://academic.oup.com/dsh> (Accessed 1 June 2024).

Google Scholar's search engine is a popular application that uses indexing infrastructures to support enquiries into authors and their publications.¹⁷⁷ Enquiry returns are based on metadata from the cataloguing of publications and other academic documents held at archives and libraries around the world (such as Dublin Core), each one tagged to a Google Scholar index by author name, thus enabling the Google search engine to restrict the number of results offered to items where metadata algorithms best satisfy the terms of the enquiry. There are other scholarly search engines, and in this thesis Google Scholar stands as a proxy for them all.¹⁷⁸

A competent researcher will soon find the locations of most of the documents whose metadata contains, for example, 'Thomas Hodgkin MD 1798 – 1866' by searching for common variations of the name and using a range of search engines. Care will need to be taken and often a judgement will be needed in making an enquiry, because (for example) Thomas Hodgkin MD was assisted in much of his public work by his similarly named nephew Thomas Hodgkin Jnr and it can be difficult to determine which Thomas Hodgkin a particular search result refers to.

If the researcher wishes to find 'Where is Thomas Hodgkin (and his name derivatives) evidenced in documents held at all archives?' this question is harder to answer, even though it is a simple and frequently asked question in the researcher's head and one that from a research perspective is quite basic to the task of researching PHL. It is also a question that is

¹⁷⁷ 'Google Scholar launched in 2004, the same year that Elsevier launched rival bibliometrics platform Scopus. Beginning in 2006, citation counts were included in search results. The scholar profiling service, "Google Scholar Citations", was launched in 2011. Citation counting drives both search and scholar profiles in different ways, and we argue that it has become a new unit of value for coordinating the scholarly economy, albeit in a manner different from JIF. While JIF quite explicitly shifted research priorities and visibility around inclusion in the SCI (De Bellis, 2014), Google Scholar is much more granular and totalizing in its shaping of research and researcher visibility according to citation count' (Goldenfein and Griffin 2022).

¹⁷⁸ British Library, National Archives, SCOPUS, COPAC, Web of Science, etc.

capable of being answered. However, currently this question cannot easily be answered online. Why not?

- Google and Google Scholar search Google's own indexes to collect results for a search enquiry. The indexes are vast and are constantly being updated by web crawlers or 'spiders'¹⁷⁹ which crawl the internet looking for items to index. The Google algorithm cannot seek for every instance of a name in every digitised document (crawler searches relate only to metadata and web page data). So it is doubtful that all of the relevant data sought will be found.¹⁸⁰
- If the data the researcher needs is in a document that has not been digitised, then neither Google search will be able to find it.
- Google Scholar is a specialist search engine and it searches only author-centred metadata (such as Dublin Core) used to compile data for search engines relating to academic publishing. It cannot search local database catalogues compiled at repositories where full-text searching technologies are often deployed. Google Scholar will find metadata for both scholarly articles and theses in which Thomas Hodgkin MD appears in the metadata information. It also uses web crawlers to enable searches for person names embedded in online texts (using full-text search capabilities).¹⁸¹

¹⁷⁹ 'Google uses crawlers and fetchers to perform actions for its products, either automatically or triggered by user request. "Crawler" (sometimes also called a "robot" or "spider") is a generic term for any program that is used to automatically discover and scan websites by following links from one web page to another. Google's main crawler is called Googlebot.' <https://developers.google.com/search/docs/crawling-indexing/overview-google-crawlers> (Accessed 14 September 2023).

¹⁸⁰ Google crawlers are continuously updated, and it is not possible for a researcher to visit all of Google's algorithms to test their utility for each enquiry.

¹⁸¹ Metadata for academic publications is produced by the publisher or archivist, and this was probably first generated by bulk copying the entire archive catalogue into metadata software such as ARK Alliance, using (for example) the Dublin Core ecology, when library catalogues began to be published online.

Some, but not all, of the publications by Thomas Hodgkin MD, and those of his contemporaries who might mention him, will have been digitised and these might be found by search engines. But a full document search for every instance of an individual name mentioned in every manuscript can only be made one global or local search engine at a time. And here, the researcher must know which search engines to use. Even for the most competent and diligent researcher, there always remains the possibility that an important Record has been overlooked. There is no universal online search engine in public use that will find references to all person names as they appear in all documents distributed across all archives. For example, nineteenth-century publications sometimes include bibliographies and indexes (they are rarely found in publications from earlier centuries) and Thomas Hodgkin MD may appear in some of them.¹⁸² For more about difficulties in online searches for Thomas Hodgkin MD, see Section 5.2.

Not all relevant books and documents have been digitised, and if they are not digitised then they cannot be searched digitally, even though the book itself may have been digitally identified through metadata (e.g. Dublin Core) and thus appear in a Google or Google Scholar search.

For the P7 Case Studies, a bespoke membership database, microfilms of society publications, manuscripts in archives and 600 genealogical searches were used to find EBPd, and none of the data (except some of the genealogical data) was findable online. The current version of the internet (Web 2.0) is of limited help to a researcher of past lives, and

¹⁸² 'Late 19th and early 20th century scholarship was dominated not by big ideas, but by methodological refinement and disciplinary consolidation. Denigrated in the later 20th century as unworthy of serious attention by scholars, the 19th and early 20th century, by contrast, took activities like philology, lexicology, and especially bibliography very seriously. Serious scholarship was concerned as much with organizing knowledge as it was with framing knowledge in an ideological construct' (Scheinfeldt 2008, cited in (Koolen, Van Gorp, and Van Ossenbruggen 2019, 380).

it is uncertain whether the Semantic Web (Web 3.0) will be of significantly more help without a system like the NAI-UID infrastructure in place to systematise EBP. Unless and until every historical document is fully digitised or there is a national infrastructure for the discovery of all person names in full-text searches across the global archive and in all researcher databases, such searches will not become routine.

3.5.3 University-level infrastructure

The quality and extent of DH Infrastructural support at the local academic level vary between academic institutions. For instance, the P7 Case Study exercises undertaken here would not have been possible without considerable support from the University of Birmingham, especially the university's Advanced Research Computing and Library Services, including the long-term support of a dedicated research software engineer (RSE). Training support was also provided online and on campus in the classroom, where DH students are routinely invited to attend training sessions primarily directed at other disciplines.

Data management from initial data collection to publication of datasets after project completion is comprehensively covered by the university's Digital Services provision and the university promotes and facilitates FAIR data collection and management requirements (see Figure 3.7).

Research Data Management (RDM)

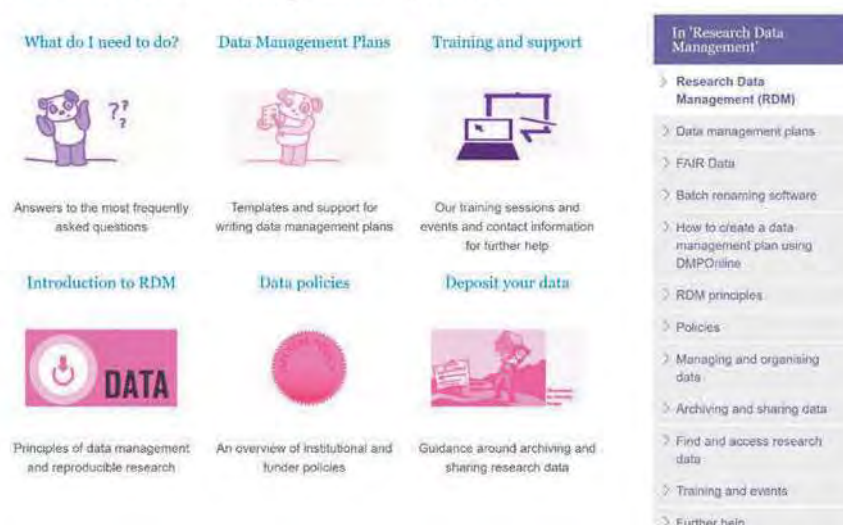


Figure 3.7 University of Birmingham RDM online infrastructure

(<https://intranet.birmingham.ac.uk/as/libraryservices/library/research/rdm/index.aspx>, Accessed 15 October 2023).

The three P7 Case Studies called for the design and build of both integrated project data flows and integrated technology pipelines. Training needs were identified early (in SQL, Python, GitHub for data management, and Gephi data visualisation graph production). Many of the skills required were learned through online learning services, for example LinkedIn, Data Camp and Stack Overflow, but early skill development and especially SQL were taught in one-to-one hourly sessions delivered by the dedicated RSE and these took place over an entire year. Local support services for digital research at the University of Birmingham were an effective single point of access to services (the university delivers some digital training services locally that are provided nationally). There was little need to look for support services outside of the university, because the breadth and depth of the university's digital support services were comprehensive and provided an effective two-way communication and learning conduit between digital support services providers and me as

researcher. Not all universities provide this level of support. The Institute of Historical Research also provided online and one-to-one support and advice, as well as offering presenting opportunities throughout the duration of the project.¹⁸³

3.5.4 Issues in infrastructure

For a critical appreciation of current DH infrastructures three influential analyses are examined here. The first is a comparative analysis, ‘Digital infrastructure for the humanities in Europe and the US: governing scholarship through coordinated tool development’ (Kaltenbrunner 2017), which highlights cultural differences between the US and the EU, explaining why each has a fundamentally different attitude to DH infrastructure. The second is a comprehensive overview of the European approach favouring top-down infrastructures: ‘Understanding the information requirements of arts and humanities scholarship’ (Benardou 2010). Finally a sceptical assessment is considered: ‘If you build it, will we come? Large scale digital infrastructures as a dead end for digital humanities’ (van Zundert 2012).

Wolfgang Kaltenbrunner sets out the cultural differences between the US and EU attitudes towards DH infrastructure.¹⁸⁴ The cultural differences can be expressed as being between (1) central coordination and community shared objectives in research in the EU and (2) an

¹⁸³ <https://www.history.ac.uk/library-digital/ihrs-digital-collections> (Accessed 11 October 2023).

¹⁸⁴ ‘European initiatives, I argue, are based on a more centralizing, technology-driven vision of digital infrastructure that serves the European Commission’s policy goal of integrating national research systems in institutional and epistemic terms. This causes a certain disconnect between tool developers and prospective scholarly users who are often unfamiliar with digital approaches, but the emphasis on central coordination also ensures that no single community gains exclusive control over technology development. In the US, by contrast, the original impetus to adopt a concerted strategy for digital infrastructure has not been provided by science policy makers and administrators, but by researchers in the area of DH. These scholars have successfully promoted a sociotechnical view of infrastructure as an emergent, evolutionary phenomenon, which also implies that conceptual and managerial authority should be situated at well-established DH centers’ (Kaltenbrunner 2017, 275).

evolutionary approach where research hubs themselves determine their individual scholastic traditions in the US. What is common to both cultures is the focus on researchers and their tools (rather than focusing on data).¹⁸⁵

This observation is important for this thesis, which concerns EBPB and the NAI system which might manifest differently in the two systems, where the location of authority and control matters more than research practices or the choice of research tools.¹⁸⁶ In summary, US infrastructures are more likely to support different, independently structured research hubs, and an evolutionary approach to research, whereas the EU infrastructures support a more collaborative and centralised approach.

Regarding the EU approach to DH infrastructure, Kaltenbrunner notes that this is enshrined at the highest level in the ERA, and then cascades down to DARIAH and CLARIN for example.¹⁸⁷ He asserts that 'A striking contrast to the European case is that the NEH's infrastructure policy is not codified in any comparable level of detail'.¹⁸⁸

¹⁸⁵ 'In theoretical terms, the vast majority of social scientists have adopted the highly influential framework proposed by Star and Ruhleder (1996). Infrastructure here is conceptualized not as a specific thing, but as a delicate ecology of interrelated socio-technical practices of different user groups (Edwards et al. 2007, 2009; Ribes and Lee 2010). All of these lines of research have in common that they tend to focus on the interaction of researchers with digital technology in the context of particular projects, studied through ethnographic or interview-based methods. Usually, they adopt a constructivist perspective in the sense of stressing the mutual shaping of infrastructure and research practices' (Kaltenbrunner 2017, 277).

¹⁸⁶ 'Current infrastructure policies constitute attempts at longer-term strategic planning in which the dispersed research instruments used in particular areas of study are subjected to an encompassing organizational and administrative framework ... the authority to interpret what type of infrastructure is needed, and who should have control over its development and maintenance, becomes a crucial topic of analysis' (Kaltenbrunner 2017).

¹⁸⁷ 'This vision of infrastructure, with its firm belief in the epistemic benefits of technologically mediated collaboration, as well as its strong emphasis on coordinated development, is informed by a specific policy strategy of the EC. For more than a decade, European policy makers have pursued the strategic goal of creating an integrated European Research Area (ERA). Their normative assumption is that the continent's scientific and economic competitiveness could be vastly improved if the European research landscape were transformed from a patchwork of national research systems with relatively isolated institutional and disciplinary structures into a more homogeneous whole' (Kaltenbrunner 2017, 284).

¹⁸⁸ '[P]roviding support to a conceptually proactive DH community, rather than trying to steer them in a topdown fashion: "Cyberinfrastructure can't be built alone. It is important that the NEH speaks with the community on a regular basis to ensure our funding strategies are best suited to help the field (Smith 2009)." A

Kaltenbrunner's conclusion to his study of the US cultural approach to DH infrastructure mirrors the cynical approach of Van Zundert, although Kaltenbrunner does recognise the efficiencies of the EU approach. He also notes that these benefits and efficiencies will only be realised if EU researchers go on to fully participate *en masse* in the EU scheme.¹⁸⁹

Agiatis Benardou provides an exemplar of a call for global infrastructures to support DH researchers in line with the EU cultural model for DH infrastructure.¹⁹⁰ She discusses a substantial exercise that analysed researcher activities in DH,¹⁹¹ which she describes as 'a required step for the evidence-based evaluation and definition of functional specifications of the planned digital research infrastructure for arts and humanities research, conducted as

striking contrast to the European case is that the NEH's infrastructure policy is not codified in any comparable level of detail' (Kaltenbrunner 2017, 296).

¹⁸⁹ '[I]t is still unclear to what extent digital infrastructure will actually be taken up across the humanities at large. Especially the European approach of developing a suite of tools that serves a large bandwidth of academics, often inexperienced in digital scholarship, is particularly likely to create friction with local disciplinary practices (Borgman 2009; Collins et al. 2012; Fry and Talja 2007). There is thus a chance that continued disinterest in digital methods on the part of "traditional" scholars will ultimately make the use of digital infrastructure as a regulatory technology unviable' (Kaltenbrunner 2017, 303).

¹⁹⁰ 'All in all, our initial analysis so far indicates that arts and humanities scholars engage in, and value highly, not only information seeking activities, but also research activities related to the curation of information objects such as primary and secondary data, and epistemic objects; arts and humanities researchers, in that sense, are curators par excellence of scholarly information, playing a key part in transforming "raw" (primary) into "institutional" (secondary) facts (Searle, 1997), through augmenting information objects semantically through annotation and edition, and through transforming them into knowledge objects by means of scholarly writing and publication. This conclusion, if confirmed, may have important repercussions on the specification of e-infrastructures, and also on our understanding of digital curation process with regard to research resources in the arts and humanities' (Benardou 2010, 28).

¹⁹¹ 'This paper reports on research of scholarly research practices and requirements conducted in the context of the Preparing DARIAH European e-Infrastructures project, with a view to ensuring current and future fitness for purpose of the planned digital infrastructure, services and tools. It summarises the findings of earlier research, primarily from the field of human information behaviour as applied in scholarly work, it presents a conceptual perspective informed by cultural historical activity theory, it introduces briefly a formal conceptual model for scholarly research activity compliant with CIDOC CRM, it describes the plan of work and methodology of an empirical research project based on open-questionnaire interviews with arts and humanities researchers, and presents illustrative examples of segmentation, tagging and initial conceptual analysis of the empirical evidence. Finally, it presents plans for future work, consisting, firstly, of a comprehensive re-analysis of interview segments within the framework of the scholarly research activity model, and, secondly, of the integration of this analysis with the extended digital curation process model we presented in earlier work' (Benardou 2010, Abstract).

part of the Preparing DARIAH European e-Infrastructures project' (Benardou 2010, 27). The report includes a flow chart (Figure 3.8).¹⁹²

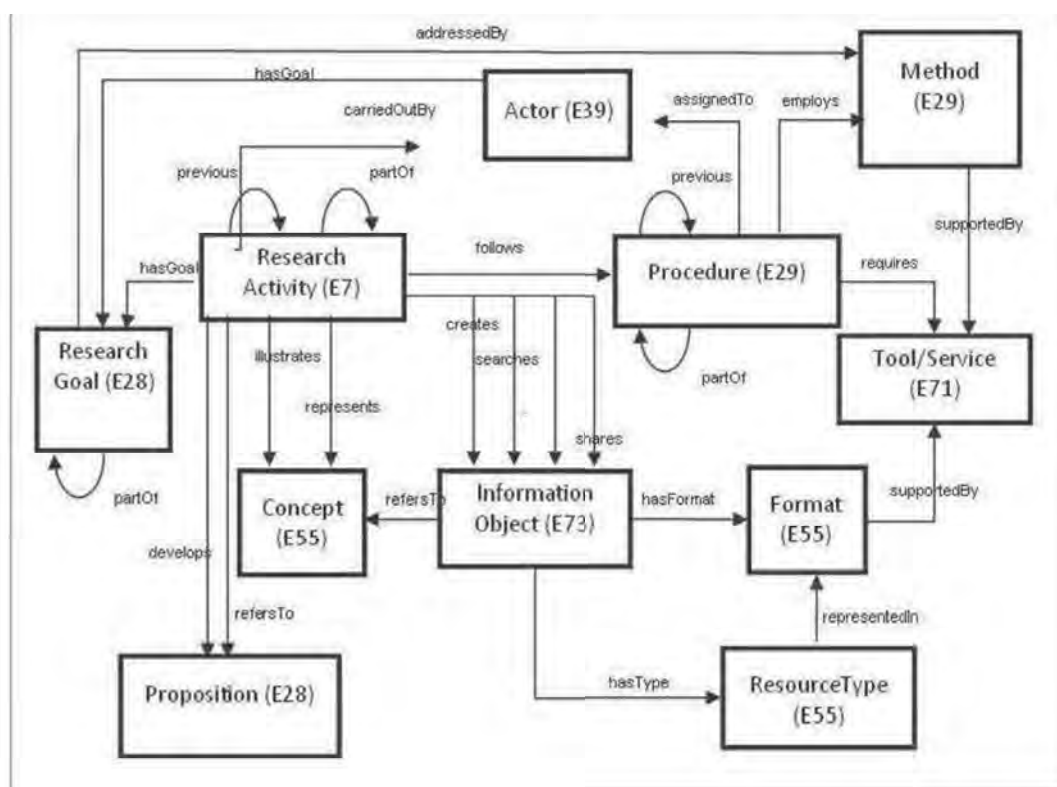


Figure 3.8 'Understanding the information requirements of arts and humanities scholarship' (Benardou 2010, 23).¹⁹³

¹⁹² 'This paper presents our approach to understanding, eliciting and analysing user requirements for information in scholarly research, a required step for the evidence-based evaluation and definition of functional specifications of the planned digital research infrastructure for arts and humanities research, conducted as part of the Preparing DARIAH European e-Infrastructures project. It summarises the findings of earlier research, primarily from the field of human information behaviour as applied in research and scholarly work, it presents a conceptual perspective informed by cultural historical activity theory, it introduces briefly our conceptual model for scholarly research activity (which constitutes the first concrete output of this research), it describes the plan of work and methodology of the empirical research project, and presents illustrative examples of segmentation, tagging and initial conceptual analysis of the empirical evidence' (Benardou 2010, 19).

¹⁹³ See also (Benardou et al. 2010, No page numbers): 'The entity Research Activity is the basic construct for representing research processes. Being a subclass of the CIDOC CRM E7 Activity, this entity is endowed with all the properties describing E2 Temporal Entity, E4 Period and E5 Event (successive classes on the hierarchy path above E7) in addition to those of E7. Of particular interest in our case are: P4 (has time-span), P119 (meets in time with), P7 (took place at), P9 (consists of), P11 (had participant), P14 (carried out by), P16 (used specific object), P17 (was motivated by), P20 (had specific purpose), P21 (had general purpose), P125 (used object of

Benardou's schema is similar to many of the schemas produced by other EU-based DH analysts, in that it envisages one closed system that embraces all research, with directional (or sequential) decision-making flows, rather like the process flow diagrams common in industry to model industrial processes.¹⁹⁴ In this way, in the EU academic sphere, serious efforts have been made to understand DH as a closed system, and closed systemic thinking of this sort has supported the development of much pan-national and community-wide infrastructure building throughout the EU and its neighbour states, as explained by Benardou.¹⁹⁵ She is here considering a 'one size fits all' approach to DH infrastructure provision and lauds the EU approach in its assessment and planning decision-making processes, as the conclusion to her report shows.¹⁹⁶

Joris van Zundert, although writing five years before Kaltenbrunner and two years after Benardou, encapsulates a concern that the European focus on large-scale infrastructures in DH (if it is focused solely on the researcher and the researcher's tools, as this thesis has

type), P134 (was continued by). These are not shown in Figure.1 for the sake of clarity. Nevertheless, the properties previous (sub-property of P134) and partOf (subproperty of P9) of Research Activity are shown in order to stress their prominent role in defining structure. This entity can be used for the documentation of accomplished as well as planned research processes through respective subclasses. Comparisons of corresponding property values allow inferences on the applicability of procedures and the actual use of resources', (Benardou 2010, 23).

¹⁹⁴ <https://www.sciencedirect.com/topics/engineering/process-flow-diagram> (Accessed 10 October 2023).

¹⁹⁵ 'Our objective is, thus, to establish a conceptually sound, pertinent with regard to actual scholarly practice, and elegant model of scholarly research activity, encompassing both "object" (structure) and "process/practice" (functional) perspectives, and amenable to operationalisation as a tool for:

- structuring and analysing the outcomes of evidence-based research on scholarly practice and requirements, and
- producing clear and pertinent information requirements, and specifications of architecture, tools and services for scholarly research in a digital environment' (Benardou et al. 2010).

¹⁹⁶ '[T]he model is meant to act as a descriptive framework for better discovery, summarisation and understanding of relationships between specific scholarly activities, research goals, information objects, methods, and tools/services at the instance level. It may, therefore, be useful as a conceptual structure – or information architecture – for better communication among stakeholders (such as policy makers, archivists, repository managers, technologists and scholars) and institutions involved in the specification of requirements and affordances of digital repositories, services and tools intended to support scholarly research work' (Benardou et al. 2010, No page numbers).

shown) will not be a sufficiently useful aid to small-scale researchers (it may even be seen as an unnecessary burden by Independent Researchers). Van Zundert thinks it is perhaps towards small-scale researchers that the democratising ambitions of the digital world should turn and so it is there that attention should be focused.¹⁹⁷

Van Zundert recognises a dilemma in large-scale disciplinisation in DH if the dominant research model in fact turns out to be small scale (similar to the Independent Researcher model proposed here).¹⁹⁸ He finds that current research interests in DH are universally wide and exploratory, innovative and unpredictable, and that DH research is likely to be undertaken either in academic research hubs or by small-scale research teams or even Independent Historians. He goes on to say, ‘modelling highly specific distributed web services is a more promising avenue for sustainability of highly heterogeneous humanities digital data than standards enforcement and current encoding practices’ (van Zundert 2012, 165).¹⁹⁹

Van Zundert is not opposed to some level of infrastructure provision; he recognises that initiatives that concern ‘hosting and safeguarding research data’ are appropriate and should

¹⁹⁷ ‘The necessary generalizations and standardizations, management, and development processes that large infrastructures need to apply to cater to wholesale humanities are at odds with well-known aspects of innovation. Moreover, such generalizations close off many possibilities for exploring new modeling and computing approaches. I argue that methodological innovation and advancing the modeling of humanities data and heuristics is better served by flexible small-scale research focused development practices’ (van Zundert 2012, 165).

¹⁹⁸ ‘Agile software development works in short bursts of creativity, called iterations or sprints, which can be as short as a week, or even less, but never more than 3 weeks. A sprint begins with a discussion between researcher and developers on what needs to be developed; it ends with the evaluation of the results by the same researcher and developers. The next iteration is planned as an answer to the changes in thinking that the experience provoked in both the researcher and the developers. In this way the actual tool or software evolves ever more into what the particular researcher actually needs, and not what some design committee thinks might be needed by all researchers’ (van Zundert 2012, 176).

¹⁹⁹ He notes too: ‘[T]his is not a problem as long as large digital infrastructures are aimed only at hosting and safeguarding research data. Given machine-negotiable access services to that data, then any tool of any kind might be applied. Tools that are relatively easy to generalize (concordancing services for instance) could be maintained more stably on such an infrastructure too. Yet this would still allow for “agile development space” to add and use tools on less institutionalized infrastructure’ (van Zundert 2012, 177).

be encouraged. He recommends that they can support and assist in innovative, short-term research needs in what he calls an ‘agile development space’ as long as such developments do not stifle innovation in creative research.²⁰⁰ He concludes that ‘There can be no absolutism in standards of conformance if we value open research practices’ (van Zundert 2012, 174).²⁰¹

Since the middle of the twentieth century considerable work in disciplinisation has taken place in DH in both the US and the EU, for instance at DARIAH²⁰² and NeMo,²⁰³ both of which provide academic-level infrastructure including proposed universal ontologies and

²⁰⁰ ‘It is nearly impossible to establish what a generalized infrastructure would look like for high-end innovative projects geared towards humanities research – the sorts that involve experimental pattern detection, large scale analysis of noisy data, and exploratory knowledge visualizations. This near-impossibility follows from the experimental character of the research. The uncertain and volatile nature of innovation determines that it is hard to establish the forms and requirements of any underlying technology or infrastructure’ (van Zundert 2012, 169).

²⁰¹ Moreover, ‘Coding and modeling are more than just collateral of the academic activities within DH; they are central to the whole enterprise. If we shift our central focus here, and take the infrastructure itself as less central, we will create the right context for truly groundbreaking engagement with humanities research data in virtual environments. What we do not need is precisely the bulky concrete highways; we can make do with the landscape that is already taking shape out there. Some bricks, mortar, shovels and gravel would be nice though, as well as a manual on how to use them’ (van Zundert 2012, 184).

²⁰² ‘DARIAH’s mission is to empower research communities with digital methods to create, connect and share knowledge about culture and society. We work towards developing an infrastructure that supports researchers working in the diverse community of practice known as the arts and humanities to build, analyse and interpret digital or hybrid resources. As such, DARIAH supports and enhances the sustainable development of digitally-enabled research and teaching through its network of people, knowledge, content, methods and tools. Our main areas of activities aim towards ensuring that humanities researchers are:

- able to assess the impact of technology on their work in an informed manner,
- access the data, tools, services, knowledge, and networks they need seamlessly and in contextually rich virtual and human environments
- produce excellent, digitally-enabled scholarship that is reusable, visible and sustainable.’

<https://www.dariah.eu/about/mission-vision> (Accessed 15 October 2023). See also (Anderson, Blanke, and Dunn 2010).

²⁰³ ‘NeMO NeDiMAH Methods Ontology. The NeDiMAH Methods Ontology (NeMO) is a comprehensive ontological model of scholarly practice in the arts and humanities, the development of which is undertaken through the ESF Research Network NeDiMAH NeMO is a CIDOC CRM-compliant ontology which explicitly addresses the interplay of factors of agency (actors and goals), process (activities and methods) and resources (information resources, tools, concepts) manifest in the scholarly process. It builds on the results of extensive empirical studies and modelling of scholarly practices performed by the Digital Curation Unit in projects DARIAH and EHRI. NeMO incorporates existing relevant taxonomies of scholarly methods and tools, such as TaDIRAH, the arts-humanities.net and Oxford taxonomies of ICT methods, DHCommons, CCC-IULA-UPF and DiRT, through appropriate mappings of the concepts defined therein onto a semantic backbone of NeMO concepts. It thus enables combining documentary elements on scholarly practices of different perspectives and using different vocabularies.’ <http://nemo.dcu.gr/index.php> (Accessed 15 October 2023).

taxonomies in DH. Although there is still much work to be done,²⁰⁴ much has already been achieved in the disciplinisation of DH in both geographical spheres.²⁰⁵ But, as Van Zundert warns, the eventual utility of building compliance-based infrastructure too rigidly at the global level is in doubt.

Kaltenbrunner, Benardou and Van Zundert show that the development of DH infrastructure is considerable in both the US and Europe, but that the approach of the two schools is radically different, with the US following an evolutionary approach supporting DH research in decentralised hubs supported by both national and sponsor funding, and the EU adopting a centralised, collaborative and somewhat proscriptive approach. (van Zundert 2012) provides a warning in that the EU approach depends on wide and extensive support (buy-in) from the research community or it may fail. His warning is a serious one given the relative immaturity of DH as a discipline and the 'Big Tent' nature of its research communities. A major concern for this thesis is the almost universal assumption in both geographies that

²⁰⁴ 'Consensus-based ontologies (in history, music, archaeology, architecture, literature, etc.) will be necessary, in a computational medium, if we hope to be able to travel across the borders of particular collections, institutions, languages, nations, in order to exchange ideas. Those ontologies will in turn exist in a network of topics, a web of "trading zones", to use a term that Willard McCarthy has used to explain humanities computing' (Unsworth 2002, no page numbers).

²⁰⁵ 'The pragmatic response is that DH is a discipline because it has the characteristics of one. Its scholarly societies include the European Association for Digital Humanities (which grew out of ALLC) and the Alliance of Digital Humanities Organizations (ADHO). The latter was founded c.2002 and is an umbrella organisation that includes new and more established members such as the ACH and scholarly societies that represent the interests of DH communities beyond Europe and North America, namely in Japan, Canada, and Australasia. The field's first journal *CHum* was founded in 1966. Today, its leading international journals include *DSH: Digital Scholarship in the Humanities* (founded by the ALLC in 1986 as *Literary and Linguistic Computing*) and *Digital Humanities Quarterly*, published by ADHO and founded in 2007 by Julia Flanders. Journals with a more regional focus also exist, for example, *Digital Studies / Le champ numérique*, founded in 1992 and published by the Societe canadienne des humanités numériques. Numerous monographs, edited collections, and the field's first Reader (M. Terras et al. 2013) have been published on the subject in the past years. DH's first major conference is usually said to have been held in Yorktown Heights in 1964 and sponsored by IBM (see Bessinger and Parrish 1965). Today, its major conference is held annually: more than 750 delegates attended Digital Humanities 2014 in Switzerland, where the acceptance rate was approximately 30 %, roughly equivalent to some leading Computer Science conferences. At present c.200 DH centres exist worldwide (according to CentreNet); ... in 2011, 134 different academic courses worldwide offering DH were identified and anecdotally it is clear that still more have since joined those ranks. It is more common for DH teaching programmes to be embedded in existing departments, for example, in University College London the DH MA/MSc is offered by the Department of Information Studies' (Nyhan and Flinn 2016, 7).

infrastructure needs in DH only mean supporting archives and researchers and their tools.

This thesis calls for a renewed focus on the study, nurture and codification of information on PHL – what it is, where it is and how it can be digitised as representative data and structured through an NAI. Only in this way will metadata allow information contained in Records to be findable, accessible, interoperable and reusable – fulfilling FAIR principles.

3.6 Research technologies: the database

Marijn Koolen, Jasmijn Van Gorp and Jacco Van Ossenbruggen provide a helpful definition of ‘digital tools’ used in a research context, placing emphasis on features such as the use of multiple tools, tools used for data transformations and the wide range of research activities performed using digital tools.²⁰⁶ They also recognise the complexity and difficulties that digital tools present in use, compared to the (perhaps over-simplified) model of the pre-digital historian seated in an archive and leafing through physical sources.²⁰⁷

Koolen, Van Gorp and Van Ossenbruggen identify three phases of digital research activity: finding, modelling and reporting. The first phase begins when the researcher initially accesses physical and/or digital archives to explore Record collections, either whole or in part (typically using a combination of online search engines and local physical or digital

²⁰⁶ ‘An increasing set of digital tools has been developed with which digital sources can be selected, analysed, and presented. Many tools go beyond key word search and perform different types of analysis, aggregation, mapping, and linking of data selections, which transforms materials and creates new perspectives, thereby changing the way scholars interact with and perceive their materials’ (Koolen, Van Gorp, and Van Ossenbruggen 2019, 368).

²⁰⁷ ‘Moreover, many digital tools allow scholars to transform, aggregate, count, classify, link, and visualize the underlying data. With these modeling steps, they further change the materials they are studying. There is as yet little common understanding within and across humanities disciplines of how these steps affect the relation between research questions and materials and how these activities differ from traditional practice in terms of interpreting and contextualizing digital data’ (Koolen, Van Gorp, and Van Ossenbruggen 2019, 369).

archival catalogues), and then later explores specific documents using full-text searches or reading at a physical archive.²⁰⁸ The finding phase culminates when the digital historian uses digital tools to interrogate (in whole or in part) the curated subset of digital items collected. This finding phase activity needs careful consideration, especially when the data examined purports to be a direct or a disseminated referent to a physical Record. Digital enquiries on collection catalogues are usually based on metadata which will have probably been ‘cleaned’ and ‘codified’ (for example into a Dublin Core datafile format tagged with the key words the archivist chose) when created. In the second phase digital data is usually ‘modelled’ to best fit the research criteria of the historical enquiry, although this can easily be influenced by the peculiarities of the specific archival digital collection-based and archive-based digitising systems.²⁰⁹

Koolen, Van Gorp and Van Ossenbruggen propose a schematic (Figure 3.9) and a critique (Table 3.2) to include in the first phase of research that gives importance to tool criticism. Tool criticism (they argue) must be considered with care and attention; it should effectively reproduce without diminution the discipline of source criticism traditionally used in historical research. Tool criticism must embrace both the tools chosen by the archivist in digital data creation and the tools chosen by the researcher for data modelling purposes.

²⁰⁸ ‘Key word searches are effective finding aids, but many digital archives and libraries offer additional sense-making tools to get a better understanding of what a digital corpus contains and does not contain and how it is structured, with which scholars can critically evaluate the archive as a whole. These can be indices of topics, persons or periods, faceted classifications based on various metadata fields, timeline visualizations, and documentation that provide details on selection criteria, data formats, and search functionalities’ (Koolen, Van Gorp, and Van Ossenbruggen 2019, 370).

²⁰⁹ ‘To interpret this aggregated information in a meaningful way, scholars need to consider the process by which it was generated, the selection of sources that were included or excluded in the analysis, and how the algorithm determines when chunks of data in different documents refer to the same thing. This is regardless of whether they did the aggregation themselves or used information previously aggregated by some tool. Reflecting on the choices that were made for identifying elements of interest in the data (such as topics, key words, or person names) and what alternative choices are possible can help scholars to consider how the actual choices focus the analysis on certain aspects and push others to the background’ (Koolen, Van Gorp, and Van Ossenbruggen 2019, 371).

Tool criticism must be treated with diligence if it is to fully complement source criticism and so it must be an essential part of digital project design.

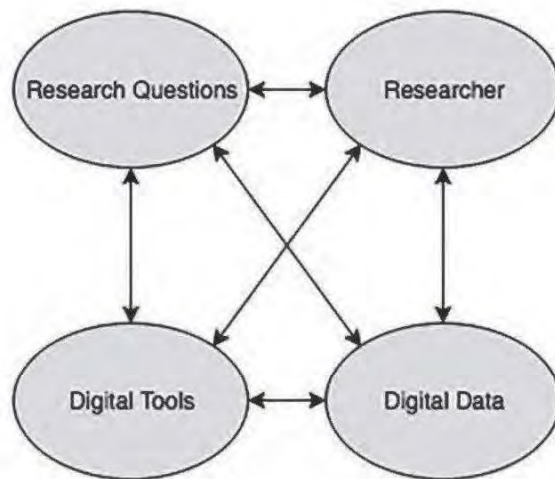


Figure 3.9 A model of interdependent concepts of digital tool criticism (Koolen, Van Gorp, and Van Ossenbruggen 2019, 373)

■ Who created the text?	■ Who made the tool?
■ What kind of document is it?	■ What kind of tool is it?
■ Where was it made and distributed?	■ When was it made?
■ When was it made?	■ Why was it made?
■ Why was it made?	■ How does the tool function?

Table 3.2 Source and tool critiques (Koolen, Van Gorp, and Van Ossenbruggen 2019, 374)

Digitisation has made access to (and use of) information contained in Records held at archives simpler and more comprehensive, but still difficulties remain and these challenge an unquestioning acceptance of and reliance on current digital ‘search and find’ systems or

the tools used to find information. Such challenges tax the researcher, who must reflect heuristically when undertaking digital research.²¹⁰ If the helpful analysis of Koolen, Van Gorp and Van Ossenbruggen is scaled up, to embrace the many and varied archival collections and also the many and varied research projects undertaken around the world, then a very complex picture emerges that current aspirations of interoperability and durability may struggle to satisfy. This is in spite of efforts to address universal digital complexity such as the promotion of FAIR principles. Technological innovation, freedom in research and the much prized enhanced creativity through digitisation will continue to test and challenge efforts to simplify and codify research into a seamless global body of work. Added to this is the fact that DH both exploits and benefits from the resources and methods developed and used in often strikingly different ways by many other research and business communities, which means that DH itself has little influence over future technological developments. This complicates the wider picture as much as it aids the DH researcher.²¹¹

²¹⁰ 'The main questions center around complex relationship between tools and data in a digital environment. The first aspect is how tools select, filter, and give access to data. Tool limitations may form a barrier to having full access to a set of data because a tool may be the only way to access them, as with Web-based tools that gives access to digital archives and heritage collections. Access to digital sources is often mediated through digital tools, which suggests an integrated criticism of tools and sources. Another issue with many digital tools working on integrated data sets is that they lack information about what data are accessible through the tool, how that data have been selected, and how tool features include or exclude certain parts of the data. This makes it hard for scholars to judge whether what they see is all there is, or that other data have been filtered out or is simply not available in the tool' (Koolen, Van Gorp, and Van Ossenbruggen 2019, 381).

²¹¹ 'The surge of research projects engaged in digitizing historical sources led to the creation of large digital collections, which have since been integrated into research practices and are now essential for the making of new historical knowledge. Against this background, highly structured and painstakingly curated collections of data have emerged, and at the same time, various digital tools and methods have been conceived, developed, and applied in order to analyze these collections. The question of whether these changes give rise to new research questions and approaches remains open, although recently it has been the focus of increased scholarly attention' (Siebold and Valleriani 2022, 171).

3.6.1 EBPD databases

At the heart of the digital study of EBPD are data tables, databases and the human relationships discovered in them. Figure 3.10 is an illustration of the relationships between persons connected to Joseph Prestwich, John Evans and John Lubbock, whom Clive Gamble in *Making deep history: zeal, perseverance, and the time revolution of 1859* argued were leading figures of the ‘time revolution’ of 1859 (when geological time began to assert itself over biblical time) (Gamble 2021).

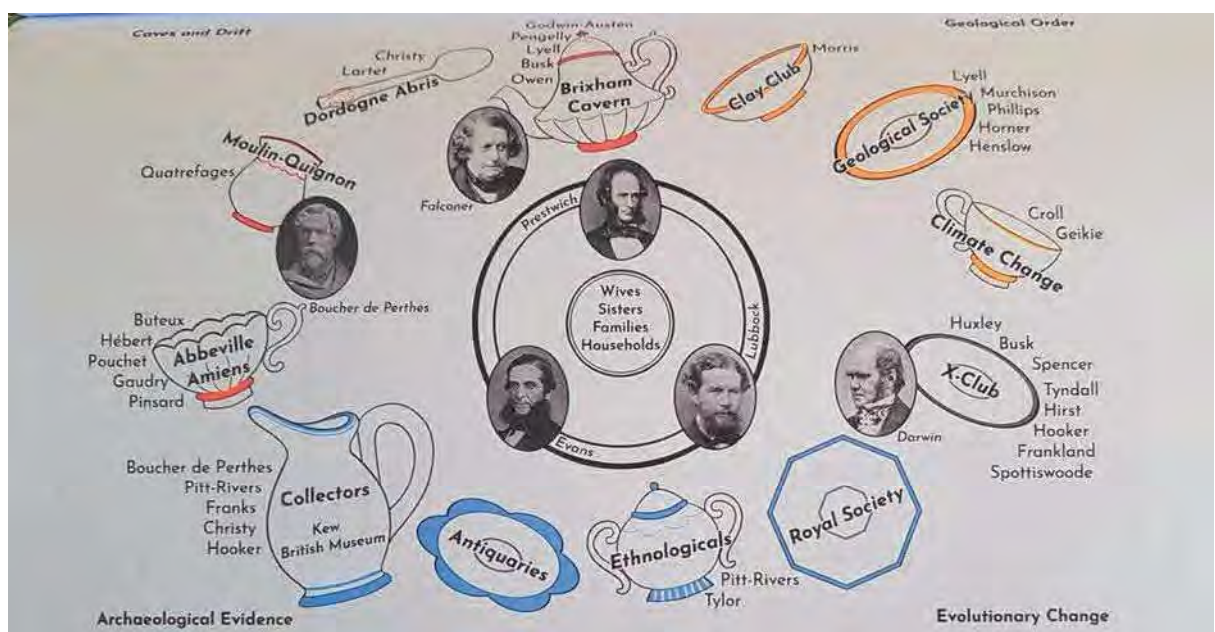


Figure 3.10 ‘The spinning plates of family, friendship and acquaintance that circled the time revolution. The result was a unification of knowledge where natural and human history were drawn together at different timescales.’ Original artwork Kaylea Raczkowski-Wood (Gamble 2021, 19)

This illustration may seem crude to technological eyes, but nevertheless it is a subtle and useful illustration of the relationships over time that Gamble explores in his study. A more technical representation of the relationships between connected persons might look like Figure 3.11. This is an Entity Relationship Diagram (ERD), the structure that organises EBPB records, and it is common to most relational databases.

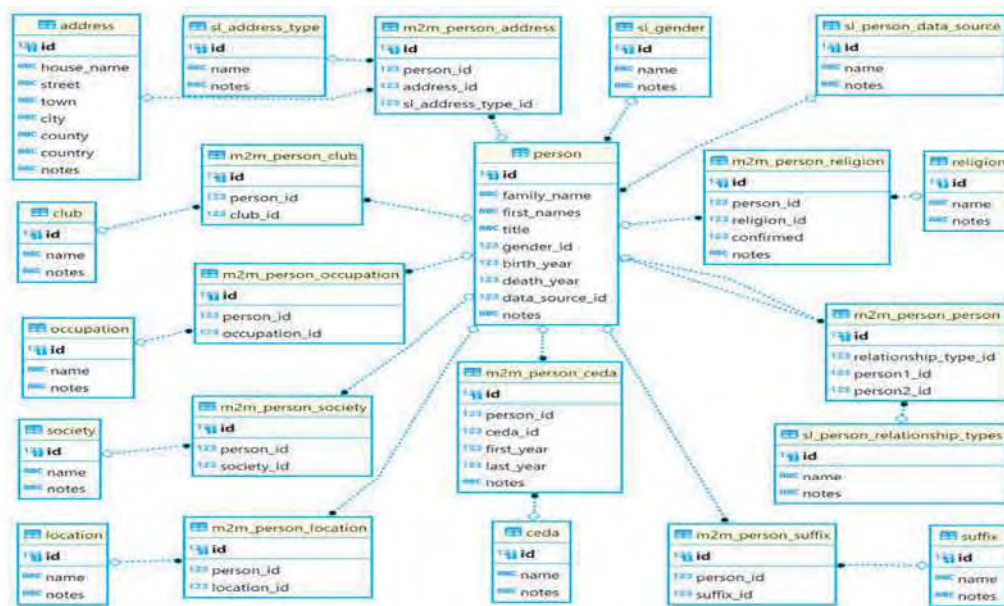


Figure 3.11 The author's schema of his CEDA digital database

The difference between the two representations is that one is a freely drawn illustration and the other a digital ERD schema. Gamble's illustration is a diagram (artwork) and it therefore stands alone, its utility being in the image alone, whereas the author's schema is digital, open-source, electronically interoperable and reproducible, and it therefore meets FAIR principles. What is most important is that it is attached to the open-source dataset it represents so that the data it references is easily found.

The use of digital tools is one of the main characteristics of the digital turn in DH, as Tom Scheinfeldt recognises,²¹² but it has been shown here that sometimes the prevalent focus on tools at the expense of data masks the fact that without good data, the very best tools are of questionable use.

3.6.2 Relational databases

Databases have a long history and the central concept of the database, the data table with its rows of records and columns of attributes in which DH information is digitally ordered and classified, has a deep structure.²¹³ The concept of the table prefigures digital databases and is a fundamental concept in the organisation of human thought, as Figure 3.12 illustrates.

²¹² 'We are entering a new phase of scholarship that will be dominated not by ideas, but once again by organizing activities, both in terms of organizing knowledge and organizing ourselves and our work. My difficulty in answering the question "What's the big idea in history right now?" stems from the fact that, as a digital historian, I traffic much less in new theories than in new methods. The new technology of the Internet has shifted the work of a rapidly growing number of scholars away from thinking big thoughts to forging new tools, methods, materials, techniques, and modes of work which will enable us to harness the still unwieldy, but obviously game-changing, information technologies now sitting on our desktops and in our pockets. These concerns touch all scholars' (Scheinfeldt 2008).

²¹³ 'Databases are an ubiquitous feature of life in the modern age, and yet the most all encompassing definition of the term "database" – a system that allows for the efficient storage and retrieval of information – would seem to belie that modernity. The design of such systems has been a mainstay of humanistic endeavor for centuries; the seeds of the modern computerized database being fully evident in the many text-based taxonomies and indexing systems which have been developed since the Middle Ages. Whenever humanists have amassed enough information to make retrieval (or comprehensive understanding) cumbersome, technologists of whatever epoch have sought to put forth ideas about how to represent that information in some more tractable form' (Ramsay 2004, 177).

The Table of CASUALTIES.

The Years of our Lord	1647	1648	1649	1650	1651	1652	1653	1654	1655	1656	1657	1658	1659	1660	1661	1662	1663	1664	1665	1666	1667	1668	1669	1670	1671	1672	1673	1674	1675	1676	1677	1678	1679	1680	1681	1682	1683	1684	1685	1686	1687	1688	1689	1690	1691	1692	1693	1694	1695	1696	1697	1698	1699	1700	1701	1702	1703	1704	1705	1706	1707	1708	1709	1710	1711	1712	1713	1714	1715	1716	1717	1718	1719	1720	1721	1722	1723	1724	1725	1726	1727	1728	1729	1730	1731	1732	1733	1734	1735	1736	1737	1738	1739	1740	1741	1742	1743	1744	1745	1746	1747	1748	1749	1750	1751	1752	1753	1754	1755	1756	1757	1758	1759	1760	1761	1762	1763	1764	1765	1766	1767	1768	1769	1770	1771	1772	1773	1774	1775	1776	1777	1778	1779	1780	1781	1782	1783	1784	1785	1786	1787	1788	1789	1790	1791	1792	1793	1794	1795	1796	1797	1798	1799	1800	1801	1802	1803	1804	1805	1806	1807	1808	1809	1810	1811	1812	1813	1814	1815	1816	1817	1818	1819	1820	1821	1822	1823	1824	1825	1826	1827	1828	1829	1830	1831	1832	1833	1834	1835	1836	1837	1838	1839	1840	1841	1842	1843	1844	1845	1846	1847	1848	1849	1850	1851	1852	1853	1854	1855	1856	1857	1858	1859	1860	1861	1862	1863	1864	1865	1866	1867	1868	1869	1870	1871	1872	1873	1874	1875	1876	1877	1878	1879	1880	1881	1882	1883	1884	1885	1886	1887	1888	1889	1890	1891	1892	1893	1894	1895	1896	1897	1898	1899	1900	1901	1902	1903	1904	1905	1906	1907	1908	1909	1910	1911	1912	1913	1914	1915	1916	1917	1918	1919	1920	1921	1922	1923	1924	1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937	1938	1939	1940	1941	1942	1943	1944	1945	1946	1947	1948	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035	2036	2037	2038	2039	2040	2041	2042	2043	2044	2045	2046	2047	2048	2049	2050	2051	2052	2053	2054	2055	2056	2057	2058	2059	2060	2061	2062	2063	2064	2065	2066	2067	2068	2069	2070	2071	2072	2073	2074	2075	2076	2077	2078	2079	2080	2081	2082	2083	2084	2085	2086	2087	2088	2089	2090	2091	2092	2093	2094	2095	2096	2097	2098	2099	2100	2101	2102	2103	2104	2105	2106	2107	2108	2109	2110	2111	2112	2113	2114	2115	2116	2117	2118	2119	2120	2121	2122	2123	2124	2125	2126	2127	2128	2129	2130	2131	2132	2133	2134	2135	2136	2137	2138	2139	2140	2141	2142	2143	2144	2145	2146	2147	2148	2149	2150	2151	2152	2153	2154	2155	2156	2157	2158	2159	2160	2161	2162	2163	2164	2165	2166	2167	2168	2169	2170	2171	2172	2173	2174	2175	2176	2177	2178	2179	2180	2181	2182	2183	2184	2185	2186	2187	2188	2189	2190	2191	2192	2193	2194	2195	2196	2197	2198	2199	2200	2201	2202	2203	2204	2205	2206	2207	2208	2209	2210	2211	2212	2213	2214	2215	2216	2217	2218	2219	2220	2221	2222	2223	2224	2225	2226	2227	2228	2229	2230	2231	2232	2233	2234	2235	2236	2237	2238	2239	2240	2241	2242	2243	2244	2245	2246	2247	2248	2249	2250	2251	2252	2253	2254	2255	2256	2257	2258	2259	2260	2261	2262	2263	2264	2265	2266	2267	2268	2269	2270	2271	2272	2273	2274	2275	2276	2277	2278	2279	2280	2281	2282	2283	2284	2285	2286	2287	2288	2289	2290	2291	2292	2293	2294	2295	2296	2297	2298	2299	2300	2301	2302	2303	2304	2305	2306	2307	2308	2309	2310	2311	2312	2313	2314	2315	2316	2317	2318	2319	2320	2321	2322	2323	2324	2325	2326	2327	2328	2329	2330	2331	2332	2333	2334	2335	2336	2337	2338	2339	2340	2341	2342	2343	2344	2345	2346	2347	2348	2349	2350	2351	2352	2353	2354	2355	2356	2357	2358	2359	2360	2361	2362	2363	2364	2365	2366	2367	2368	2369	2370	2371	2372	2373	2374	2375	2376	2377	2378	2379	2380	2381	2382	2383	2384	2385	2386	2387	2388	2389	2390	2391	2392	2393	2394	2395	2396	2397	2398	2399	2400	2401	2402	2403	2404	2405	2406	2407	2408	2409	2410	2411	2412	2413	2414	2415	2416	2417	2418	2419	2420	2421	2422	2423	2424	2425	2426	2427	2428	2429	2430	2431	2432	2433	2434	2435	2436	2437	2438	2439	2440	2441	2442	2443	2444	2445	2446	2447	2448	2449	2450	2451	2452	2453	2454	2455	2456	2457	2458	2459	2460	2461	2462	2463	2464	2465	2466	2467	2468	2469	2470	2471	2472	2473	2474	2475	2476	2477	2478	2479	2480	2481	2482	2483	2484	2485	2486	2487	2488	2489	2490	2491	2492	2493	2494	2495	2496	2497	2498	2499	2500	2501	2502	2503	2504	2505	2506	2507	2508	2509	2510	2511	2512	2513	2514	2515	2516	2517	2518	2519	2520	2521	2522	2523	2524	2525	2526	2527	2528	2529	2530	2531	2532	2533	2534	2535	2536	2537	2538	2539	2540	2541	2542	2543	2544	2545	2546	2547	2548	2549	2550	2551	2552	2553	2554	2555	2556	2557	2558	2559	2560	2561	2562	2563	2564	2565	2566	2567	2568	2569	2570	2571	2572	2573	2574	2575	2576	2577	2578	2579	2580	2581	2582	2583	2584	2585	2586	2587	2588	2589	2590	2591	2592	2593	2594	2595	2596	2597	2598	2599	2600	2601	2602	2603	2604	2605	2606	2607	2608	2609	2610	2611	2612	2613	2614	2615	2616	2617	2618	2619	2620	2621	2622	2623	2624	2625	2626	2627	2628	2629	2630	2631	2632	2633	2634	2635	2636	2637	2638	2639	2640	2641	2642	2643	2644	2645	2646	2647	2648	2649	2650	2651	2652	2653	2654	2655	2656	2657	2658	2659	2660	2661	2662	2663	2664	2665	2666	2667	2668	2669	2670	2671	2672	2673	2674	2675	2676	2677	2678	2679	2680	2681	2682	2683	2684	2685	2686	2687	2688	2689	2690	2691	2692	2693	2694	2695	2696	2697	2698	2699	2700	2701	2702	2703	2704	2705	2706	2707	2708	2709	2710	2711	2712	2713	2714	2715	2716	2717	2718	2719	2720	2721	2722	2723	2724	2725	2726	2727	2728	2729	2730	2731	2732	2733	2734	2735	2736	2737	2738	2739	2740	2741	2742	2743	2744	2745	2746	2747	2748	2749	2750	2751	2752	2753	2754	2755	2756	2757	2758	2759	2760	2761	2762	2763	2764	2765	2766	2767	2768	2769	2770	2771	2772	2773	2774	2775	2776	2777	2778	2779	2780	2781	2782	2783	2784	2785	2786	2787	2788	2789	2790	2791	2792	2793	2794	2795	2796	2797	2798	2799	2800	2801	2802	2803	2804	2805	2806	2807	2808	2809	2810	2811	2812	2813	2814	2815	2816	2817	2818	2819	2820	2821	2822	2823	2824	2825	2826	2827	2828	2829	2830	2831	2832	2833	2834	2835	2836	2837	2838	2839	2840	2841	2842	2843	2844	2845	2846	2847	2848	2849	2850	2851	2852	2853	2854	2855	2856	2857	2858	2859	2860	2861	2862	2863	2864	2865	2866	2867	2868	2869	2870	2871	2872	2873	2874	2875	2876	2877	2878	2879	2880	2881	2882	2883	2884	2885	2886	2887	2888	2889	2890	2891	2892	2893	2894	2895	2896	2897	2898	2899	2900	2901	2902	2903	2904	2905	2906	2907	2908	2909	2910	2911	2912	2913	2914	2915	2916	2917	2918	2919	2920	2921	2922	2923	2924	2925	2926	2927	2928	2929	2930	2931	2932	2933	2934	2935	2936	2937	2938	2939	2940	2941	2942	2943	2944	2945	2946	2947	2948	2949	2950	2951	2952	2953	2954	2955	2956	2957	2958	2959	2960	2961	2962	2963	2964	2965	2966	2967	2968	2969	2970	2971	2972	2973	2974	2975	2976	2977	2978	2979	2980	2981	2982	2983	2984	2985	2986	2987	2988	2989	2990	2991	2992	2993	2994	2995	2996	2997	2998	2999	3000	3001	3002	3003	3004	3005	3006	3007	3008	3009	3010	3011	3012	3013	3014	3015
-----------------------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------



Figure 3.13 Image of the room of the Enlightenment, British Museum

(<https://www.britishmuseum.org/collection/galleries/enlightenment>, Accessed 1 November 2023)

The modern digital database reproduces this common, deep-rooted urge to organise and classify things in ordered rows and columns – in tables. Berg, Seymour and Goel provide a helpful timeline for the history of digital databases starting in the mid-1960s (see Table 3.3).²¹⁵

²¹⁵ See also (Ryan, Emerson, and Robertson 2014, 125 - 130).

<ul style="list-style-type: none"> • ‘Direct-access storage (disks and drums)’ ²¹⁶
<ul style="list-style-type: none"> • Followed by the ‘relational database model [which] was conceived by E.F. Codd in 1970’²¹⁷
<ul style="list-style-type: none"> • In the 1980s, ‘Structured Query Language [(SQL)] became the intergalactic standard’²¹⁸
<ul style="list-style-type: none"> • In 1991 the advent of the ‘World Wide Web’ begins and ‘Microsoft Access’ is created in 1992
<ul style="list-style-type: none"> • Also, around this time, ‘open source’ solutions came online with widespread use of GCC (GNU Compiler Collection), CGI (Computer Generated Imagery), Apache, and MySQL.
<ul style="list-style-type: none"> • Finally in 1997 XML emerges ²¹⁹

Table 3.3 Timeline of the history of digital databases ((Berg, Seymour, and Goel 2013, 30-33))

²¹⁶ <https://www.ibm.com/docs/en/aix/7.2?topic=subsystem-direct-access-storage-devices-dasds> (Accessed 2 November 2023).

²¹⁷ ‘E. F. Codd first proposed the relational model in a 1970 article in Communications of the ACM entitled “A Relational Model of Data for Large Shared Databanks.” Codd’s proposal endeavored to overcome the limitations of previous systems, which had suffered from difficulties related both to inefficient (which is to say slow) access and unwieldy storage mechanisms – inefficiencies that often resulted from redundancies in the underlying data representation. Codd’s model made great strides forward in both areas, and yet his achievement is perhaps more acutely evident in the mathematical presentation of his ideas. One researcher, who refers to the 1970 paper as “probably the most famous paper in the entire history of database management,” notes: “It was Codd’s very great insight that a database could be thought of as a set of relations, that a relation in turn could be thought of as a set of propositions ..., and hence that all of the apparatus of formal logic could be directly applied to the problem of database access and related problems”. This fundamental idea has spawned a vast literature devoted to database theory, and while there have been several major additions to the relational model, the relational databases of today continue to operate on the basis of Codd’s insights’ (Ramsay 2004, 178).

²¹⁸ ‘SQL stands for Structured Query Language. SQL lets you access and manipulate databases. SQL became a standard of the American National Standards Institute (ANSI) in 1986, and of the International Organization for Standardization (ISO) in 1987.’ https://www.w3schools.com/sql/sql_intro.asp (Accessed 31 October 2023).

²¹⁹ https://www.w3schools.com/xml/xml_what_is.asp (Accessed 2 November 2023).

The digital relational database first emerged in 1970, and since then it has grown and developed largely in terms of processing power and ease of use, but the underlying concept and basic structure of the data table remain constant. Writing in 2013, Berg predicted the next steps in the development of distributed databases, distributed data and processing, and he did so with remarkable accuracy.²²⁰

It can be concluded that relational databases are today widely used, that they are sometimes open source and that they are a mature and globally distributed technology. They are ideally suited for the purpose of EBPd and research, where organised ‘person with attributes’ and relationships are the research objectives. Databases in one form or another are likely to remain part of the deep structure of the digital world, and therefore will remain a core DH tool for some time.²²¹

²²⁰ ‘Currently databases are beginning to take on even more complex logic. Another feature that is being expanded is the concept of separation of location from the abstract concept of the database itself; which Codd defined long ago. This feature enables a database to reside in more than one location and to be queried as a continuous unit. Such instances are called distributed or federated databases. A portion of a database can be in New York and another in Boston and a query requested to count all the customers would then be run simultaneously at both locations. This is also made possible due to the increase in the speed of networks’ (Berg, Seymour, and Goel 2013, 34).

²²¹ ‘Initially, DH research projects used database technologies developed in the context of other disciplines and applications, mostly around the intersection of computer science and business. Among the most widely used technology is the relational database, which was developed in the early 1970s and released as a product in 1978. Compared to earlier database systems (which were based on hierarchical or network database models), the relational database simplified the management, processing, and querying of data. This is partly due to the introduction of the table as an organizing principle, which enables data to be organized into rows and columns that can be related to each other. For historical research, the relational database meant that historians could conduct qualitative research digitally’ (Kemman 2021, quoted in (Siebold and Valleriani 2022, 171).

3.6.3 Graph databases

Anna Siebold and Matteo Valleriani, in ‘Digital perspectives in history’, provide a definition of the graph database.²²² Michael Friendly’s ‘Milestones Project’,²²³ a comprehensive web-based long timeline series analysis of the development of data visualisation technologies, makes clear that data visualising systems have been in common use for centuries.²²⁴ After listing several recent specialist partial histories of data visualisation, Friendly asserts that current researchers are often unaware of the history of data visualisation.²²⁵ Friendly traces the history of data visualisation at least to the year 1644 (see Figure 3.14).

²²² ‘Network analyses (similar to graph databases) emphasize the relationships between people, places, events, objects, or concepts. They aim to describe the character of a network, its density or central orientation, the nature of relationships in the network, and who or what occupies a central role. They allow, in contrast to a single document or biography, for the description of complex behavior in a network of relationships over time’ (Siebold and Valleriani 2022, 172).

²²³ ‘Milestones in the history of thematic cartography, statistical graphics, and data visualization’, Michael Friendly and Daniel J. Denis. <https://www.datavis.ca/milestones/index.php?page=introduction> (Accessed 30 October 2023).

²²⁴ ‘The earliest seeds arose in geometric diagrams and in the making of maps to aid in navigation and exploration. By the 16th century, techniques and instruments for precise observation and measurement of physical quantities were well-developed – the beginnings of the husbandry of visualization. The 17th century saw great new growth in theory and the dawn of practice – the rise of analytic geometry, theories of errors of measurement, the birth of probability theory, and the beginnings of demographic statistics and “political arithmetic”. Over the 18th and 19th centuries, numbers pertaining to people-social, moral, medical, and economic statistics began to be gathered in large and periodic series; moreover, the usefulness of these bodies of data for planning, for governmental response, and as a subject worth of study in its own right, began to be recognized.’ <https://www.datavis.ca/milestones/index.php?page=introduction> (Accessed 28 October 2023).

²²⁵ ‘But there are no accounts that span the entire development of visual thinking and the visual representation of data, and which collate the contributions of disparate disciplines. In as much as their histories are intertwined, so too should be any telling of the development of data visualization. Another reason for interweaving these accounts is that practitioners in these fields today tend to be highly specialized, often unaware of related developments in areas outside their domain, much less their history’ (Friendly 2008a, 2).

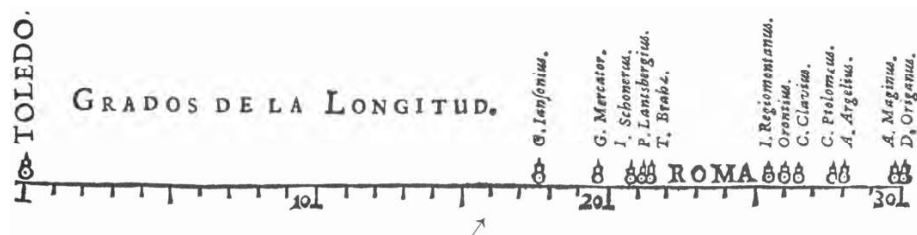


Figure 3.14 Langren's 1644 graph of determinations of the distance, in longitude, from Toledo to Rome. Taken from Tufte 1997, 15; (Friendly 2008a, 4).²²⁶

Many of the common graphical display layouts used today in digital technologies also have a very long history. (Friendly 2008a, 4) illustrates this by saying, 'William Playfair (1759–1823) is widely considered the inventor of most of the graphical forms widely used today—first the line graph and bar chart (Playfair, 1786), later the pie chart and circle graph (Playfair, 1801)'. He calls the first half of the nineteenth century 'The Golden Age'.²²⁷ Thereafter, Friendly claims, a 'dark age' exists until around 1950. He suggest that this was due to a shift between 1900 and 1950 towards a preference for precise statistical calculations, which only accurate numbers can convey (for example when calculating numbers to several decimal places); data visualisations are unsuited to handling numbers to high levels of precision.²²⁸

²²⁶ See also (Friendly 2008a, 3) 'What is notable is that van Langren could have presented this information in various tables—ordered by author to show provenance, by date to show priority, or by distance. However, only a graph shows the wide variation in the estimates; note that the range of values covers nearly half the length of the scale. Van Langren took as his overall summary the center of the range, where there happened to be a large enough gap for him to inscribe "ROMA." Unfortunately, all of the estimates were biased upwards; the true distance (16 ± 300) is shown by the arrow.'

²²⁷ 'By the mid-1800s, all the conditions for the rapid growth of visualization had been established. Official state statistical offices were established throughout Europe e, in recognition of the growing importance of numerical information for social planning, industrialization, commerce, and transportation. Statistical theory, initiated by Gauss and Laplace, and extended to the social realm by Quetelet, provided the means to make sense of large bodies of data. What started as the Age of Enthusiasm (Palsky, 1996) for graphics may also be called the Golden Age, with unparalleled beauty and many innovations in graphics and thematic cartography' (Friendly 2008b, 5).

²²⁸ 'There were few graphical innovations, and, by the mid-1930s, the enthusiasm for visualization which characterized the late 1800s had been supplanted by the rise of quantification and formal, often statistical, models in the social sciences. Numbers, parameter estimates, and, especially, standard errors were precise. Pictures were—well, just pictures: pretty or evocative, perhaps, but incapable of stating a "fact" to three or more decimals. Or so it seemed to statisticians' (Friendly 2008a, 6).

Since 1950, with the emergence of the computer as a research tool, digital visualisation has developed into the technology with which we are familiar today. Friendly cites three important influences on this final phase of development: (1) John W. Tukey (Exploratory Data Analysis), (2) Jacques Bertin (*Sémiologie Graphique*), and (3) the rise of computer processing.²²⁹

Siebold and Valleriani point out that while there does not appear to be ‘an all-encompassing universal network theory’ in DH from an analytics perspective,²³⁰ borrowing technologies from the sciences developed for more complex analysis has appeal because the needs of DH researchers are easily met. DH datasets tend to be smaller and therefore they do not challenge the limits of big data concepts as they now arise in scientific disciplines. Graph database technologies such as the Gephi affordance used in the P7 project (see Section 6.19) exploit relationships between entities such as persons and their familial relationships. They are an essential part of the EBP system and the Semantic Web.

²²⁹ ‘(a) In the USA, John W. Tukey began the invention of a wide variety of new, simple, and effective graphic displays, under the rubric of “Exploratory Data Analysis.” (b) In France, Jacques Bertin published the monumental *Sémiologie Graphique* (Bertin, 1967, 1983). To some, this appeared to do for graphics what Mendeleev had done for the organization of the chemical elements, that is, to organize the visual and perceptual elements of graphics according to the data. (c) Finally, computer processing of data had begun, and offered the possibility to construct old and new graphic forms by computer programs. True high-resolution graphics were developed, but would take a while to enter common use’ (Friendly 2008a, 7).

²³⁰ ‘It is important to point out that there is no all-encompassing universal network theory. What does exist, however, is a shared core of analytical concepts, such as density, centrality, and community building. Many of these have been transformed into indicators, implemented in software, and are presented in textbooks (Lemerrier 2015). Although initially scholars primarily invoked Social Network Analysis (SNA)—the approach that seemed most appropriate to their subject matters—recent developments reveal the limitations of SNA, which is why some historians are moving toward approaches that were originally developed in the frame of physics of complex systems. The reason for this shift lies in the nature of the datasets that the historical sources generate; they are usually somewhat smaller than those of the natural or life sciences’ (Siebold and Valleriani 2022, 173).

3.7 Historians/data scientists

Reflection on the utility and performance of the P7 project allows revisiting of the dual-project model for DH research collaboration proposed by (Breure, Doorn, and Boonstra 2006) (Figure 3.15): the historian and the data scientist.

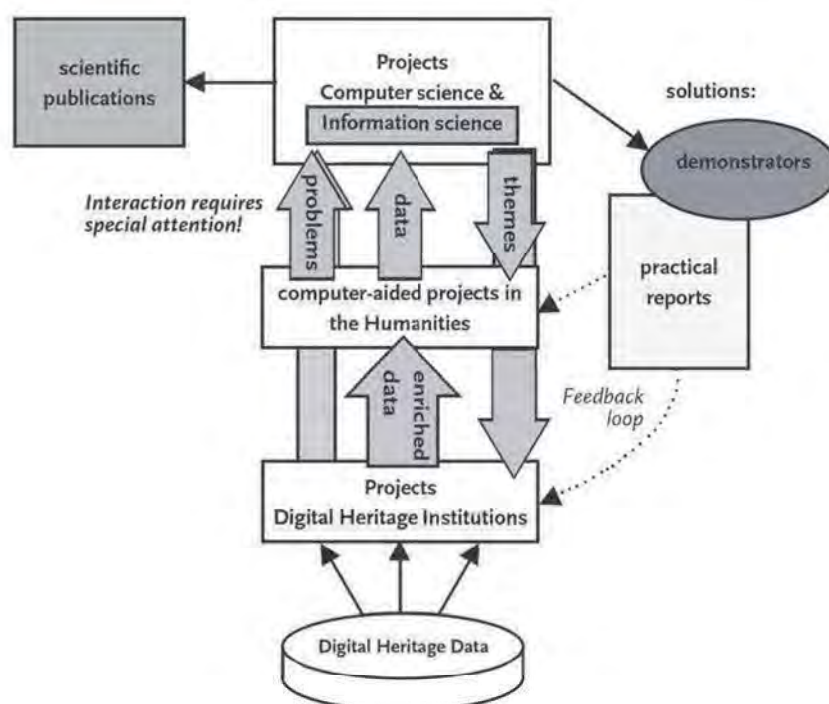


Figure 3.15 Recommended dual-project model for collaboration (Breure, Doorn, and Boonstra 2006, 98)

Breure, Doorn and Boonstra's dual-project model can be modified by defining 'Digital Heritage Data' as that sourced from genealogy, archives and other sources, and including in their schema 'genealogical institutions' to include 'Digital Heritage Institutions', and finally their 'computer-aided projects in the humanities' can be redefined as a 'multi-skilled research unit'. What then emerges is a simple but highly reproducible RSE-Independent Researcher model. This model is still a compact 'nuclear' mode of humanistic and

technological working practice and because of this it is durable, dynamic and reproducible (see Figure 3.16).

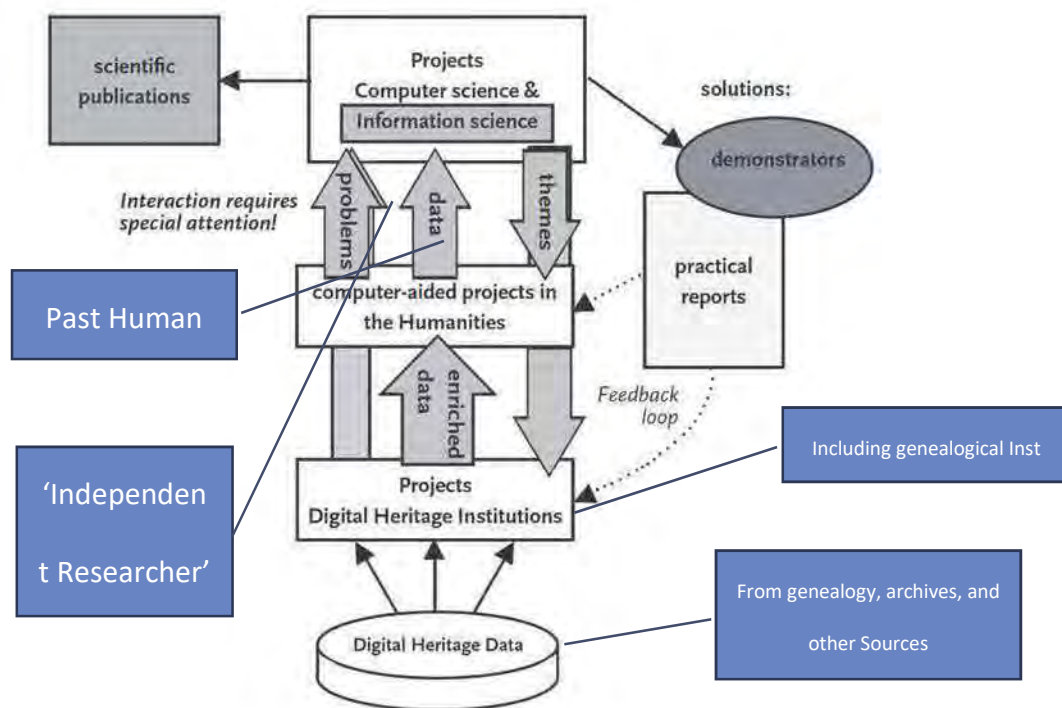


Figure 3.16 The Breure, Doorn and Boonstra model modified by Kelvin Beer-Jones

If the EBPD approach to the comprehensive mapping of nineteenth-century biographical data were to be widely promoted and adopted, it would greatly improve the efficiency of future archival research by making available to later researchers a bedrock of open-source, irreducible person names and biographical attributes (found once, used many times). Therefore, the EBPD system has the potential to eliminate needless repetition in future researcher searches, and in the long term considerably to extend the web of digitally captured, valuable, biographical EBP with attributes, until the study of PHL as a discipline emerges – even if it does so slowly and sporadically, one group exercise at a time. Nonetheless, this system will gradually and cumulatively achieve mass data mapping as

many research projects spontaneously and dynamically are formed and work together, as long as the EBPd evidence collected is then preserved and made available to other researchers, in open-source and traversable forms, from one location to another. It is not necessary that one grand global database be built, only that human names and biographical attributes discovered in instances of EBPd in Records conform to EBPd best practice and that datasets are held where they can be easily found. To be able to do this, EBPd copied to and held at repositories must be searchable and traversable, so that when needed by other researchers new discrete sets of EBPd can be compiled by copying EBPd already found and organised without losing the path back to Records. To function in a disciplined and methodologically sound way, each research project must therefore commit to following appropriate standards when discovering and offering up at archive new EBPd (model standards are largely already largely in place in DH and other related academic disciplines),²³¹ and indeed also in genealogy.²³²

EBPd and biographical and other person attributes uncovered in research in the future under the EBP system will be structured and organised, allowing researchers to be able to respond to the challenges and opportunities of working in a developed and extensive digital world. However, this will only occur if those finding and using EBPd in their own research accept the task of recording it and placing it in local digital repositories in a structured and codified way, referencing data both to the Record of its source and the NAI index of the life referenced. It is highly likely then that the quality of EBPd taken up by researchers and genealogists will be relatively high, perhaps almost as high as that of archivists working with their digital affordances.

²³¹ <https://www.dariah.eu/activities/working-groups/guidelines-and-standards/> (Accessed 1 November 2023).

²³² https://www.familysearch.org/en/wiki/Genealogical_Proof_Standard (Accessed 15 October 2023).

Many Independent Historians and genealogists, working separately, each on their own individual research projects, could nonetheless, through their research activities, carefully extract, codify and locate at repositories vast amounts of EBP. They would do so if, during their individual research activities, they encountered and brought to light instances of EBP information on PHL.

The current state of the development of digitisation programmes (i.e. technological take-up, capability building and access facilitation) in DH now provides a window of opportunity for the further exploitation of information held in archives to be developed to provide academic standard EBP as a virtual global service capability. This thesis argues that by linking up genealogical and archival data, the EBP NAI system extends the scope of digital research into PHL and (given that genealogical research is a popular activity widely practised) helps to develop PHL research by enabling Independent Historians and genealogical researchers to work together using digital technologies. In this way the EBP system opens up the possibility of making new histories in new and novel ways.²³³ This thesis considers that it is Independent Historians and genealogists who will become major contributors to the study of PHL if the EBP system is adopted. This is impossible currently because many thousands of individual and poorly connected historians lack data support provision and infrastructure. To demonstrate how the EBP system would work for an Independent Historian, P7 has been undertaken and lessons from it support the arguments made here (see Chapter 6).

²³³ 'To better cope with the demands of practitioners in the digital publishing era, historians need to become more information literate, while information specialists need to better understand the specific information needs of historians. Arguably, what is needed is a new generation of practitioners who are highly trained in the craft of history as well advanced information skills, such as computer programming, database development and multimedia production' (Breure, Doorn, and Boonstra 2006, 19).

3.8 Research Software Engineers (RSEs)

RSEs are not found in every academic institution, and where they are found they tend to span academic activities across a large campus. The academic world of RSEs is perhaps better labelled ‘digital science’, of which DH is just one of many component clusters.

Whether in each RSE cluster digital science is an umbrella organisation from an RSE point of view – where the real ‘home’ of RSEs is in data science as a clearly defined grouping and from where RSEs are farmed out to other disciplines on a task-by-task basis – or data science functions as a central meeting point for RSEs whose ‘homes’ are distributed around the disciplines, varies from one academic institution to another. Whatever the organisational structure, RSE is today growing fast across academia, and that includes dedicated RSEs in DH.

Philippe et al. (2019) note: ‘The Arts and Humanities Research Council (AHRC) recently commissioned a team at the Universities of Southampton and Oxford to undertake a study into how best to support and build the skills, knowledge and capacity of the research community to utilise digital tools and infrastructure and ensure world class research.’²³⁴ It is

²³⁴ ‘Digital tools and software are revolutionising the nature of data and research across the arts and humanities community; how data is collected and analysed, and how it is managed, shared, and sustained for future generations.

> The principles of open research and the necessity of providing effective infrastructure to support the changing shape of research are driving policy and practice across UKRI, higher education and government.

> The rich disciplinary diversity that characterises arts and humanities research poses challenges for providing infrastructure and support for research skills that can meet wide ranging needs and priorities.

> There is a lack of evidence about current research practices and levels of engagement in and use of digital tools in the arts and humanities community and the implications of this for digital skills gaps and needs.

> AHRC commissioned a team at the Universities of Southampton and Oxford to undertake a study into how best to support and build the skills, knowledge and capacity of the research community to utilise digital tools and infrastructure and ensure world class research’ (Olivier et al. 2019, 9). Consider also that RSEs are not present uniformly across global human society, they are concentrated in wealthier areas. See RSEs in the world in 2018 at <https://www.software.ac.uk/blog/2018-03-12-what-do-we-know-about-rses-results-our-international-surveys> (Accessed 15 October 2023).

appropriate to discuss here that report and its findings, because the UK is one of the main centres for RSEs today. The most significant recommendation that the report makes from the point of view of this thesis is that data management and sustainability are not well provisioned.²³⁵ Selected charts from the report indicate the roles and the extent of engagement of RSEs in DH in the UK.

3.8.1 RSEs and engagement

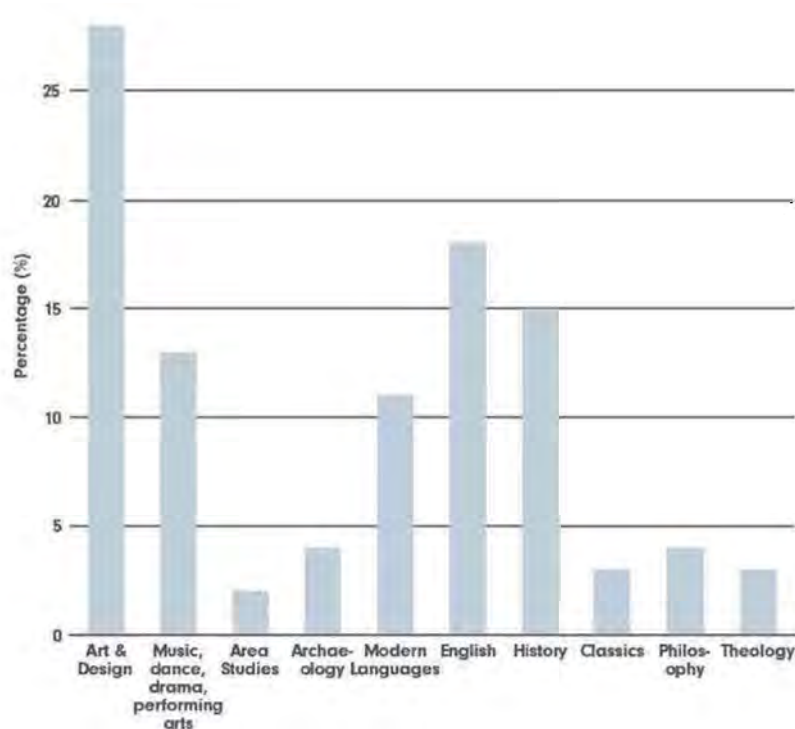


Figure 3.17 Researchers by discipline (Olivier et al. 2019, 23)

²³⁵ 'Data management and sustainability – These elements are not yet fully embedded in research processes, and there is a lack of knowledge and understanding of both the broad rationale for open research and the institutional infrastructures that support it. In addition, those infrastructures may not be functioning effectively at institutional levels.

> The AHRC should promote sustainability strategies, build these into application and review processes, audit and monitor funded research whilst also understanding the challenges researchers face at an institutional level. These processes should ensure flexibility for researchers' (Olivier et al. 2019, 60).

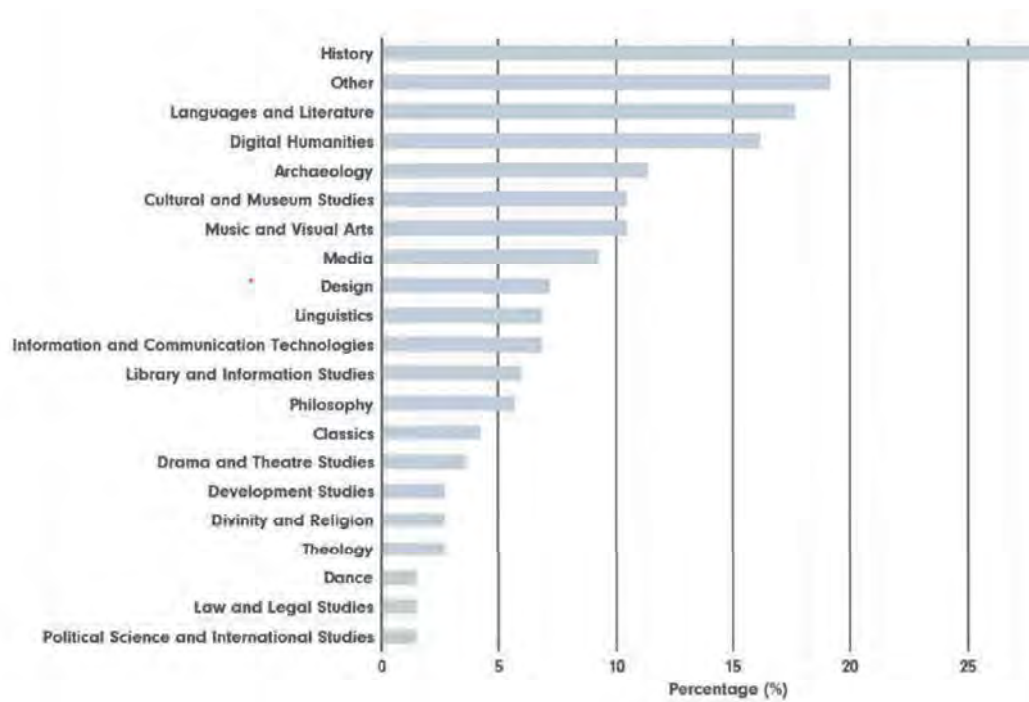


Figure 3.18 Discipline of survey respondents (Olivier et al. 2019, 27)

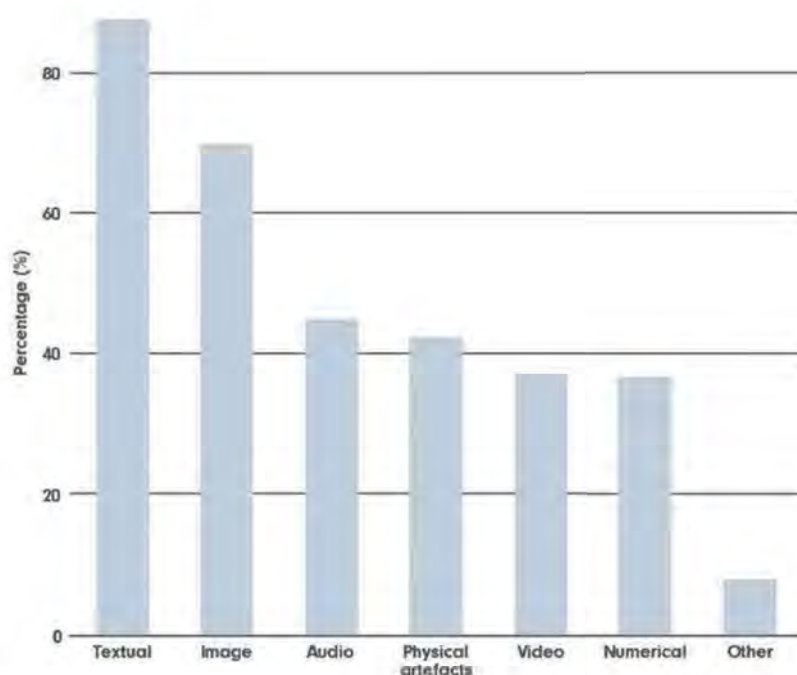


Figure 3.19 Types of research data used by arts and humanities researchers (Olivier et al. 2019, 30)

Comparing Figures 3.17, 3.18 and 3.19, we can see that around 90% of all researchers use text as data, and nearly 30% of all survey respondents were in the discipline of history, whereas history represents only 15% of humanities researchers. This indicates the intensity of DH found in the history department, where the focus is predominantly on text. There is a strong link between RSEs, textual analysis and historians in the survey and that indicates real potential support for the IRG model (which includes both historians and RSEs).

3.8.2 RSEs and data

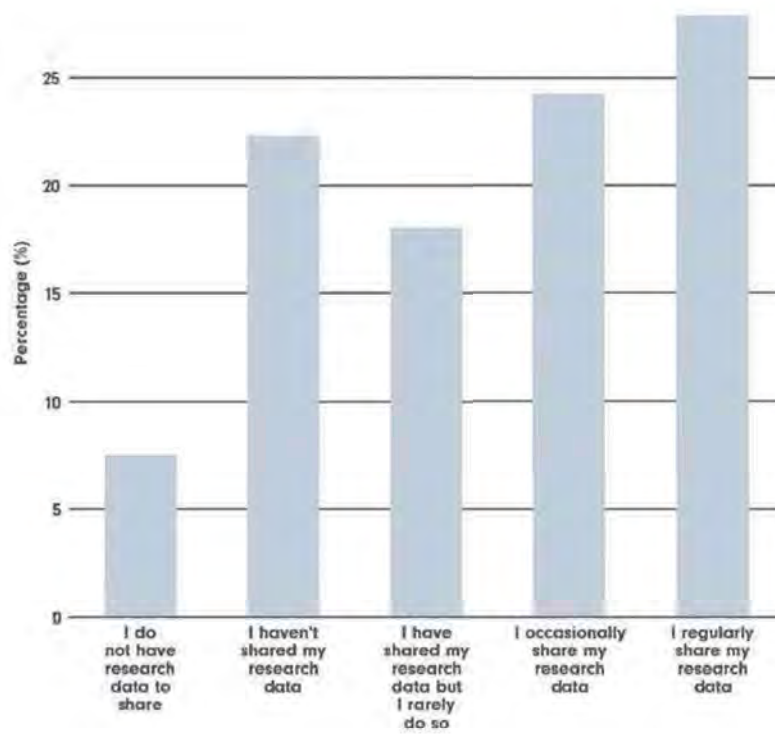


Figure 3.20 Frequency of data sharing – all respondents (Olivier et al. 2019, 36)

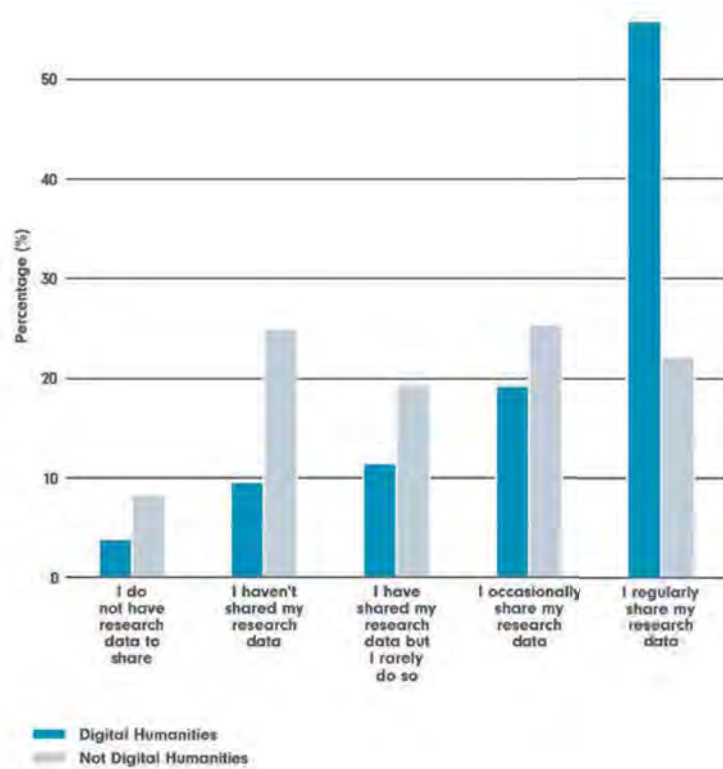


Figure 3.21 Frequency of data sharing by humanities/DH researchers (Olivier et al. 2019, 36)

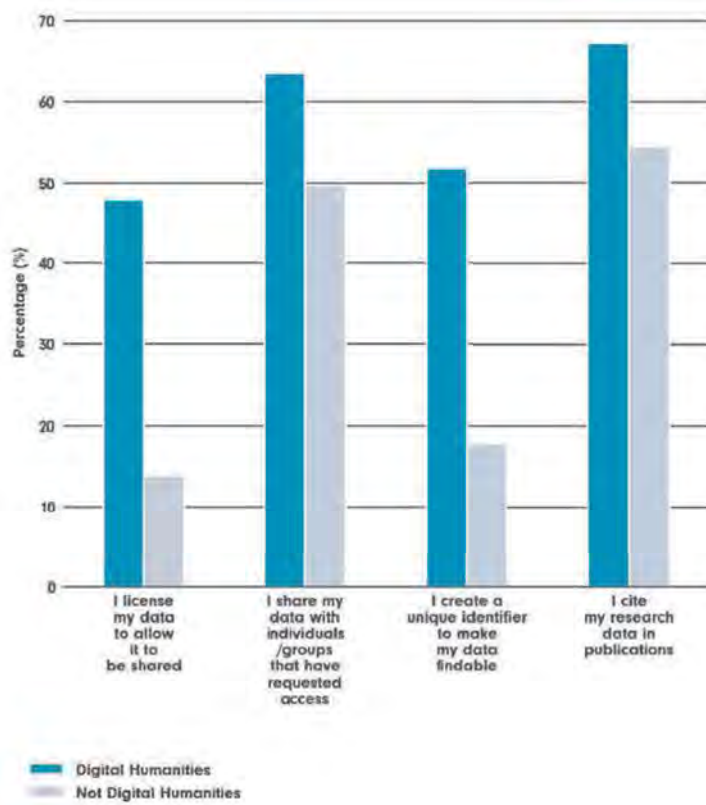


Figure 3.22 Aspects of data sharing by humanities/DH researchers (Olivier et al. 2019, 38)

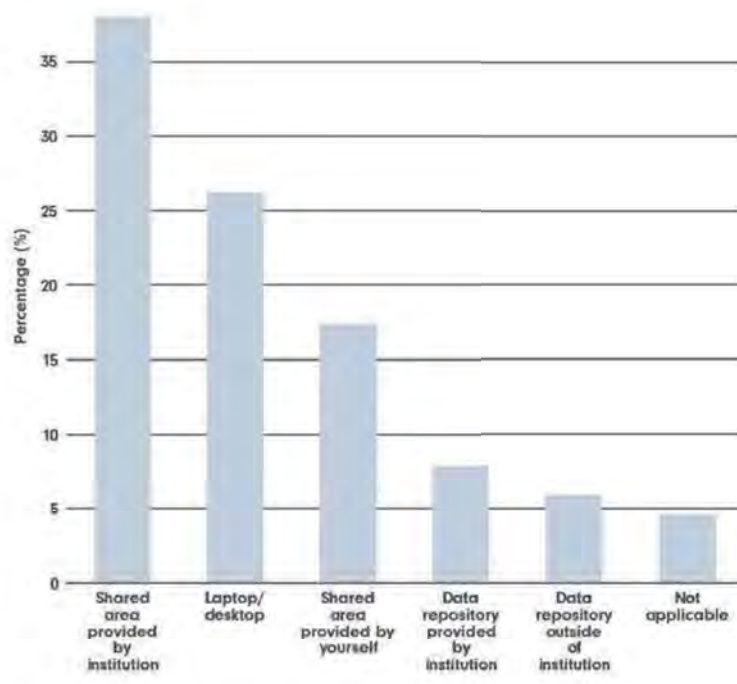


Figure 3.23 Data storage locations – all respondents (Olivier et al. 2019, 39)

Comparing Figures 3.20–3.23, we can see that of all the respondents, 28% regularly and 24% occasionally share data, but when we split out DH researchers nearly 60% regularly share data. Figure 3.22 shows that DH researchers outperform other researchers across all aspects of data sharing. In both the important aspects of licensing and the use of unique identifiers (both essential in making EBPD discoverable), DH researchers are around three times more likely to share effectively. Worryingly, although 37% of respondents store data at their home institution, more than 40% ‘store it themselves’. Unfortunately, the main report does not tell us what the scores are for DH researchers within Figure 3.23, but we can hope that storing data at institutions is more prevalent among DH researchers, as is perhaps indicated by the way in which digital researchers outperform all others across all other metrics.

Overall, this suggests that the likelihood of compliance with future EBPD and Archaeological Prosopographer standards and requirements (if they were developed) could be high. RSEs can play a critical role in promoting EBPD preservation and copying of that data to approved repositories.

For the P7 Case Study exercises, an RSE from the University of Birmingham joined this researcher and provided training to facilitate the following:

- Brainstorming the technical design of the project and its aims
- Advising on technology and data pipeline configuration
- Recommending University of Birmingham standards in research data management
- Advising on code of practice for archiving and sharing data
- Advising on open and reproducible research practices
- Database and schema design
- One-to-one training in SQLite, Visual Basic Code, DBeaver, Python Notebooks and GitHub for version control
- Problem-solving guidance

The RSE assisting a Independent Researcher concept worked well throughout. There were management issues such as cross-project communication (the RSE advised on the project first and the genealogist's data was obtained afterwards, so they never met). There were contributions that improved the work process as well as the P7 project outcomes, discussed in Chapter 6.

3.9 Chapter summary

Features such as the Scholarly Primitives and Methodological Commons highlighted by Unsworth and McCarty in 2000, and the idea of the 'Big Tent' in 2011 that data in DH is both complex and messy, set the tenor of both reflections on the past development of DH as a discipline and also what future DH infrastructure ought to be like. Another defining characteristic in the development of DH is the difference in approaches in the US and the EU, with the US stylised as a distributed system and that of the EU as centrally driven. In the UK, a hybrid system is approaching, with a mix of centralised national features alongside hub-led structures. Each culture has its strengths and weaknesses. In the US access to DH research hubs is difficult for those who live and work remote from the hubs, whereas in the EU small, innovative teams of researchers can easily be overlooked, because relatively rigid access to influence infrastructure design is limited to supporting academic institutions, which constrains spontaneity and creativity. Kaltenbrunner, Benardou and Van Zundert provided alternative analyses through which to consider the Methodological Commons and the rise of the 'Big Tent' characterisation of DH through the lens of both regional approaches. Further changes in infrastructure affordance can be anticipated if the demand for future DH services cannot be adequately met structurally by either of these major providers in the future.²³⁶ Koolen, Van Gorp and Van Ossenbruggen offered a critique of the use of research technologies in DH and the importance, durability and ubiquity of data tables and databases in DH research were explained.

²³⁶ 'Much of what is blossoming requires further elaboration and has to be translated into more widely applicable and usable tools. Better infrastructure is needed in order to guarantee a transfer of results from the methodological and technical level to the daily practice of historical research. On the contrary, denying these challenges and opportunities will, in the long run, segregate the study of history from the technical capabilities currently being developed in the information society and will turn "the computer" into an awkward tool with limited use and usability for historians' (Breure, Doorn, and Boonstra 2006, 92).

The relatively long history of DH and the now universal characteristics of the 'Methodological Commons' and the 'Big Tent' features of DH, as well as its manifestation in two different regional forms, distributed and centralised, and the enduring presence of deep structures like databases and data tables, mean that DH today is structured around universal themes, robust to a variety of organisational systems and bedded in with deep fundamental features. Before DH ossifies into hard-to-change practices, it would be timely to consider new, complementary infrastructures such as EBP and the NAI-UID, which enable research into PHL and thereby support the future of DH. The EBP system connects DH into the Semantic Web and enables all researchers to simultaneously live with and overcome some of the complexities and difficulties of 'messy data' that deserve urgent attention.

What is striking in this chapter is the absence of references to information as opposed to technologies and tools. Research into PHL is fundamentally based in the analysis of information. The only exception here is Section 3.8 which discusses genealogy. Genealogy platforms are designed to match EBP found in information as aids to researchers who are members of the general public; they are not academics. This lack of infrastructural provision for research into PHL in academia is addressed by the proposed EBP and NAI-UID system.

EBP and the NAI-UID system provide the scope for the infrastructural support that DH currently lacks. If the EBP NAI-UID system were to be adopted, it would fit well with current DH infrastructural support based around metadata and information. No doubt the early development of DH, in the area of PHL, has been based on metadata and information because these can relatively easily be represented as ordered digital data. The world of PHL is comparatively chaotic and has up to now defied digitisation. The exception is in the

digitisation of small, isolated and discrete sets of information often researched in Corpus Linguistics, information that is frequently represented as digital image files.

Chapter 4 EBP in Galleries, Libraries, Archives, Museums and Special collections (GLAMS)

Because Evidence Based prosopographical Information (EBPI) is under represented in current digitisation efforts in the GLAMS,²³⁷ this chapter examines the digital infrastructures and affordances in GLAMS in detail.

This frames a detailed analysis of both metadata systems and standards to identify where current provisions of both do not include EBP as a deliverable, but where future provision in both can embrace EBP by using the extensibility capabilities present in both metadata systems and standards.

Could digitised finding aids be a better bridge to the records and the informations contained in them that researchers are interested in? Important new developments in metadata in the areas of Universal Bibliographic Control and Archival Authority Control (AAC) in relation to Name Authority Records (NAR) show the direction of travel for standards in GLAMS relating to EBP. Finally, a detailed analysis is performed on leading affordances in GLAMS standards and name authority affordances in GLAMS. Special collections are probably not large or rich enough to develop specialist affordances at the GLAMS level, instead benefiting from the general affordances and improvements gained in host archives, so special collection do not figure greatly in this chapter.

²³⁷ 'We also discuss the significant role that structured data, much of which has been contributed by humanists employed within memory institutions and recorded in institutional information systems for the last 30 years, can potentially have in the open environment of the semantic web. These sources have been largely overlooked as a significant source for analytical humanities research, but could provide valuable and unique meaning, context and perspective, at both micro and macro levels of research' (Oldman, Doerr, and Gradmann 2015, 253).

Chapter 2 set out and described at a high level the cultural similarities and differences of three proposed levels of EBP: at national, GLAMS and research levels. Chapter 5 will explore a small representative set of the current affordances of researcher aids at the research level. There is no significant interest yet in considering how EBPI and EBP are addressed at the national level (a major recommendation of this study) and therefore there is no existing manifestation at the national level to be investigated by this thesis. However, Name Authority Record systems have a long history in GLAMS and they are the basis of the proposed NAI-UID system, so they are discussed in detail here.

This chapter is focused on GLAMS and the EBPI present in sources held in collections. The chapter discusses the current and considerable affordances of authorities, indexes, ontologies and Conceptual Reference Models, and how EBP and the proposed NAI-UID concept could fit within GLAMS and GLAMS affordances. The balance between standards development, adoption and implementation is now sufficiently established that there is a window of opportunity for the adoption of the NAI-UID system. A clear role is emerging for EBP in GLAMS. There is now considerable digital infrastructure in place through the digitisation of bibliographies, catalogues, indexes and finding aids²³⁸ and the provision of related search and research related support tools.²³⁹

²³⁸ 'Finding aid is a broad term that covers any type of description or means of reference made or received by an archival repository in the course of establishing administrative or intellectual control over archival materials. The term "finding aid" can include a variety of descriptive tools prepared by an archives (e.g., guides, calendars, inventories, box lists, indexes, etc.) or prepared by the creator of the records (e.g., registers, indexes, transfer lists, classification schemes, etc.). Such tools provide a representation of, or a means of access to, the materials being described that enables users to identify material relating to the subject of their inquiries. An archival repository's descriptive system will likely consist of various types of finding aids, each serving a particular purpose' (Society of American Archivists 2020, 58).

²³⁹ 'Libraries take a "bibliographic" approach to metadata, which is rooted in their traditional strength in describing books. Bibliographic metadata focuses on detailed descriptions of individual items that allow users to locate these items. Archives use "finding aids," descriptive inventories of collections, along with historical information necessary for understanding the material' (Riley 2017, 5).

It must be recognised, however, that while digital infrastructure design and development advance at pace, and the adoption of standards approaches universality largely through interoperability initiatives at the individual GLAMS level, take-up is still somewhat patchy and implementation of standards remains globally an enormous task. Implementation at national level still presents enormous challenges too, as the recent experience of the National Archives of France in 2021 illustrates. The challenges faced by the Records in Context (RiC; see Section 4.4.4) implementation team are set out in Table 4.1.

<ul style="list-style-type: none"> • A huge amount of heterogeneous metadata, created by generations of archivists and historians over centuries. A significant effort for digitizing them from 1990, and for updating and completing this legacy.
<ul style="list-style-type: none"> • In a series of silos, the main one contains more than 29000 archival finding aids (XML/EAD files, thus structured documents) and about 15000 authority records (XML/EAC-CPF files) on the archival creators, as well as about 20 vocabularies in a specific XML format But it is not the only one... We have a lot of databases, and other repositories (among which a digital library, and the recently implemented digital archival system).
<ul style="list-style-type: none"> • Already a lot of relations between the files in the first silo, but not really viewable and not searchable through the front-end web application (the Salle des inventaires virtuelle) and as concerns the relations between the finding aids and the authority records, of one category only (the provenance relation).
<ul style="list-style-type: none"> • The vocabularies and other authority records of the main silo can be used for indexing the finding aids, but are rather poor, not standardized and quite rarely used till now.
<ul style="list-style-type: none"> • Several end-user interfaces.
<ul style="list-style-type: none"> • Very few intuitive access points (who, when, where, what...).
<ul style="list-style-type: none"> • A lot of redundancies, from one silo to another and within the same silo (particularly the main one, where the same group of records or same record may have been described several times in several finding aids).
<ul style="list-style-type: none"> • The end user finds it difficult to understand the result lists and how the results are displayed (in the context of a finding aid, as subcomponents of it).

- | |
|--|
| <ul style="list-style-type: none">• Very few bridges to other information systems. |
|--|

Table 4.1 Implementing ICA Records in Contexts-Ontology (RiC-O) at the National Archives of France (ANF): first steps and prospects (table built from bullet points on a PowerPoint slide in (Clavaud 2021))

A detailed examination of GLAMS digital standards in authorities and related metadata developments will show that in spite of the patchiness of digitisation, (1) archival infrastructures are ready for the adoption of EBP and the NAI-UID indexing system, and (2) the adoption of EBP and the NAI-UID system will enhance current archival provision without impairing the considerable advances in digital affordances made at GLAMS over the last fifty years.

Two broad types of digitised research information provision are not of direct concern to this thesis: (1) Born Digital data and (2) academic writings such as theses, journal articles and academic papers. Nevertheless, they are briefly discussed here because their individual digitisation stories reveal and sometimes offer similarities and learnings helpful in the study of EBPI and the study of PHL, and because they are both humanities research areas that are deeply interconnected to EBPD; for example, each has its own citation and bibliographic infrastructures, and both frequently reference information sources.

This chapter is focused on the following:

- Classifications relevant to the study of EBPD in GLAMS.
- Representative leading institutions in developing authority standards in GLAMS.
- The roles of Universal Bibliographic Control (UBC) in libraries and international standards for archival authority records in establishing authority standards.

- New approaches to EBPI-related affordances in GLAMS.
- Name Authority Records (NAR) and the digital representation of ‘person names’ and other prosopographical information embedded in physical primary sources, because this is the point of interface between GLAMS metadata and EBP.

GLAMS authority systems are today both several and varied in content and structure, even after considerable recent efforts at systems simplification and standardisation. The challenges in the further development of person name authority systems remain significant and will likely persist. Paradoxically, because metadata provisioning in authority systems is not yet fixed, a window of opportunity has emerged to encourage the adoption of EBP in the future development of the NAI-UID system.

Current GLAMS digital systems have been built up over many years, often incurring massive investment in terms of money and hours worked. Digital systems improvements, while benefiting future applications, can often only accommodate rather than absorb previous digitising efforts, because the cost of restructuring or replacing past digital practices and affordances is frequently prohibitive.²⁴⁰ The ability of the digital replacements of old paper finding aids to read and work seamlessly with both existing and past digital affordances is a universal concern which drives a philosophy of caution in GLAMS digitisation. This cautious approach manifests when upgrading and sometimes replacing relatively old library bibliographic systems and archival finding aids. It is especially the case in the establishment

²⁴⁰ ‘In the early years of EAD, Tatem found that barriers included the dearth of affordable software and browsers capable of displaying EAD finding aids, and the lack of access to training in the creation of EAD-compliant finding aids. Many archives found that implementing EAD successfully meant doing significant work to “reengineer” finding aids that were incomplete or otherwise did not meet current data content standards. Many archives with limited resources struggled to hire staff with EAD expertise and establish the technological infrastructure required to create and publish EAD-encoded finding aids. In her usability studies of EAD interfaces, Yakel found that users’ lack of familiarity with the finding aid format actually deterred them from successfully navigating EAD records’ (Gracy and Lambert 2014, 102).

and management of metadata in the area of digital cataloguing and the authority indexing of NAR. GLAMS has a long tradition of competence and discipline in the development and provision of digital affordances based on metadata, such as bibliographic records in the library, finding aids in the archive, and cataloguing and index building.

This chapter demonstrates that the adoption of EBP and the NAI-UID system in GLAMS would benefit and enhance both existing and future NAR provision and make digital finding aids more useful by better connecting them to the information they discuss. It is for this reason that a representative set of prominent and recent affordances in GLAMS management and control are considered here:

- Resource Description and Access (RDA)
- International Council on Archives Records in Contexts Conceptual Model (RiC-CM)
- International Committee for Documentation of the International Council of Museums Conceptual Reference Model (CIDOC-CRM)

The discussion also shows how these could be considerably improved and extended if they adopted the NAI-UID system.

4.1 Classification of EBPI sources in GLAMS

EBPI is found in abundance in GLAMS Records and in particular among all four GLAMS general classifications: Books, Papers, Records and Objects (BPRO). Each GLAMS school has its own history of digital standards development, but taken together these have similarities (because of shared commonalities) and also important and systemic differences (because of

particular characteristics of each GLAMS area).²⁴¹ Special Collections can be found across GLAMS and sometimes have local bespoke metadata systems that are usually adaptations of earlier data management systems massaged to conform with host institution norms of practice. The area is often a subclass of host GLAMS areas, and therefore Special Collections are not usually leaders in the development of pan-GLAMS standards and authority work.²⁴²

Divisions in GLAMS between data locations and data types (documents, artefacts, images, etc.) and their overlaps when rendered as metadata make complex any analysis of the development of standards and concepts across GLAMS as a whole. For example, the management of authorities has a different manifestation in each classification, but it is also a shared concern across all of GLAMS and frequently uses commonly shared systems. Additionally, standards developed in and for individual GLAMS areas have been gradually consolidating and harmonising over the last twenty years or so while simultaneously protecting unique and important differences in the nature of each sector's collected items. As a result, different routes to digitisation have been taken in each GLAMS area, yet BPRO classification moves towards greater harmonisation of standards.²⁴³ EBP development in

²⁴¹ 'For a long time, libraries have created their own indexing tools based on the specific characteristics of collection objects, such as the physical object, its material characteristics, and its methods of production and circulation (Bianchini 2022, 64). Although most people associate archives and manuscripts in some way with libraries, the archives profession is truly distinct from the library profession, with its own established theoretical background, history, and methodology', (Morris 2009, No page numbers).

²⁴² 'Often, archivists, manuscripts curators, and/or special collections librarians work with a mixture of types of materials—institutional records or archives; manuscripts and personal papers; and rare books. The term "special collections" is generally applied to collections that include materials other than the institutional archives, such as manuscripts and rare books, although "special collections" can be used as an umbrella term to encompass these types of collections in addition to institutional archives', (Morris 2009, No page numbers).

²⁴³ '[T]here no longer exist a bibliographic, archival and museum universe that are totally separate from each other, because the distinction between what is peculiar to a field (such as a nomen) from what it is not (the associated entity, such as a person, a place, a work etc.), it allows us to see more clearly how some entities are common to all three universes and with the semantic web in general. There are, and there will always be some typical specificities of each universe, but certainly, for certain entities, these specialized universes are only one or more of the possible facets of the universe described and represented, in its totality and completeness, in the semantic web' (Bianchini 2022, 72-73).

GLAMS must therefore fit in with, and enhance, existing digitisation initiatives across all of GLAMS, if EBP is to be successfully taken up.

This chapter discusses EBPI and its treatment in GLAMS by broadly following the scope and scheme of an address to the 66th IFLA Council and General Conference in 2000 by Eeva Murtomaa, who focused on interoperability between each of the GLAMS areas.²⁴⁴ The object of this chapter is to show that, as metadata systems continue to be developed and increasingly harmonised across GLAMS, EBP and the NAI-UID system offer enhanced improvement in affordance across the entire GLAMS domain, and could become an integral part of future digitisation in GLAMS.

4.1.1 Born Digital data and academic writings – relationship to EBP

Born Digital and academic writings (theses, articles and other learned writings) are in EBP considered to be secondary sources. They therefore only have an indirect relationship to the area of focus of this thesis. Nevertheless, efforts to systematise this sector offer helpful lessons for digital development in GLAMS. Born Digital data²⁴⁵ is a relatively new field of study where the objects of research are digital items in their first instantiation, and where

²⁴⁴ 'This paper discusses interoperability between libraries, archives, and museums, focusing on how the user can have simultaneous access to all kinds of material based on a core level of description. Tools for description and conceptual data modeling are addressed, including ISAD(G) (General International Standard Archival Description), ISBD(G) (International Standard Bibliographic Description), and FRBR (Functional Requirements for Bibliographic Records). Let's do our best for finding a deeper semantic interoperability between libraries, archives and also with museums. We have to create tools, standards and interfaces to make the systems to co-operate in searching and record transferring. With international standards like the 239.50 applications we can have transparent access to wide variety of dissimilar systems e.g. of libraries and archives, even if these organisations are not using the same rules or formats internally' (Murtomaa 2000, Abstract).

²⁴⁵ <https://www.nationalarchives.gov.uk/information-management/manage-information/digital-records-transfer/what-are-born-digital-records> (Accessed 15 June 2024) and <https://www.sas.ac.uk/about-us-6/institutes-and-centres/digital-humanities-research-hub/events/born-digital-collections> (Accessed 15 June 2024).

the digitised item is also the unmediated object of research. Born Digital data brings with it its own particular concerns and among these is authentication, an interest shared with GLAMS.²⁴⁶ Equally, the management of provenance, which was identified as a major concern early in the development of Born Digital in DH, is also of concern in GLAMS. Therefore, the development of authentication systems in the Born Digital area is of interest to the development of EBP because it is a related data management concern.²⁴⁷

Anne J. Gilliland-Swetland and Philip B. Eppard in 2000 identified five features of Born Digital items that give rise to concerns about authenticity: affixedness, fixity, temporality, annotations and persons.²⁴⁸ These authenticity features of concern are also largely shared with GLAMS (and EBP), but there is one important difference between Born Digital and GLAMS and that is in the feature of affixedness. This is a lesser concern for Born Digital but a

²⁴⁶ 'With the emergence of primary Sources in digital form, the demand on the librarian and archivist to continue to support scholars by presenting them with trusted primary sources has reached a level of complexity undreamed of by the palaeographers of previous generations. The technological, ethical, conceptual, and procedural issues driving this complexity are relatively new to the humanities and information studies, and so far lack the weight of scholarly legitimacy that surrounds more traditional subdisciplines such as codicology. Even in the relatively short span of their existence, the document types and file formats present in the era of the digital archive have changed with unsettling rapidity' (Kirschenbaum et al. 2010, 32).

²⁴⁷ 'In the development of digital libraries and of digital information systems in general, increasing attention is being given to issues relating to the preservation and authenticity of digital objects in order to assure their long-term accessibility and physical and intellectual integrity [Lynch 1994, Duranti and MacNeil, 1996, Bearman and Trant, 1998, Rothenberg, 1999, Council on Library and Information Resources, 2000]. Different types of digital objects have varying preservation and authenticity requirements, however, depending upon the contexts of their creation and use. Furthermore, these requirements are also subject to differing degrees of stringency. The most basic requirements for establishing the authenticity of a digital object may be very similar to the heuristics that information literacy programs seek to inculcate in end users working with of any type of information – that is, establishing the who, what, when, where, how, and why associated with that information' (Gilliland-Swetland Anne J. 2000).

²⁴⁸ 'Fixity – Intellectual fixity is more critical than physical fixity and is generally absent, at least conceptually. How is it to be achieved? The "setting aside" of a record (e.g., through processes such as capture, registration, and storage) needs to be triggered by some intellectual event that represents the intellectual closure of that activity, or some other indication that the record has achieved the consequences it was created to achieve. Temporal Views – Can they be reconstructed? Completed records kept in live systems without being physically segregated or otherwise set aside are generally still subject to retrospective updating or reformatting when the system's data structure is changed or the system is migrated. Annotations – When annotations are made to a record after its compilation or receipt or in the course of its management, they are not readily identifiable. Juridical-Administrative Context – It is difficult to identify juridical persons involved in the creation of electronic records because they are frequently not readily visible but are inferred or implied based on the context and other intellectual elements of form in the record; are inherited values from other elements; or are inserted automatically through the presentation or display' (Gilliland-Swetland Anne J. 2000, 6/8).

critical concern for GLAMS, because without the quality of affixedness between the instantiation of information on PHL (EBPI) and its digital representation (EBPD), the connection to the primary source is broken.²⁴⁹ Considered from the perspective of GLAMS, EBPD is a digital representation of EBPI contained in a physical BPRO item. A digital representation of a physical item (particularly one that records the instantiation of a person name) should always be a fixed referent to that particular instantiation of that particular physical source, thus making affixedness a primary and critical concern for GLAMS and the EBP system.

Academic writings are secondary sources (unless they become designated primary sources) and are therefore of no direct interest in EBP. Section 3.4.2 discussed Google Scholar²⁵⁰ as an exemplar of the affordance of digital infrastructure in academic publishing through its use of the 'Alphabet' search engine deployed on a single web platform, alongside Google indexing technologies and Google finding aids. Through the combination of these, Google Scholar affords a 'publisher to platform to academic' research and academic citing facility. Chapter 3 showed that Google Scholar is a broad exemplar of a technological affordance suggesting the way in which a NAI-UID system could also be made into a web based affordance.²⁵¹ There are many platforms competing with Google Scholar, as is evident from

²⁴⁹ 'Affixedness – The notion of a record needing to be physically affixed to a medium in order to be a record (concept of the physical carrier of the record). The case study data so far indicate that the medium is incidental and transparent and does not play a significant role in assuring authenticity, except in the immediate moment of rendering the record, e.g., in a screen display' (Gilliland-Swetland Anne J. 2000, 6).

²⁵⁰ 'Google Scholar has become an important player in the scholarly economy. Whereas typical academic publishers sell bibliometrics, analytics and ranking products, Alphabet, through Google Scholar, provides "free" tools for academic search and scholarly evaluation that have made it central to academic practice. Leveraging political imperatives for open access publishing, Google Scholar has managed to intermediate data flows between researchers, research managers and repositories, and built its system of citation counting into a unit of value that coordinates the scholarly economy' (Goldenfein and Griffin 2022, Abstract).

²⁵¹ 'Rather than digesting research articles according to their semantic content as librarians had traditionally done, citation indexing enabled the organization of articles according to the works they referenced. Characterizing research according to its networks of references afforded a good proxy for content digesting,

Table 4.2, which shows the percentage of all citations found by Google Scholar compared to other main finding aids: none currently provides comparable levels of coverage or affordance, and all finding aids, including Google, have their weaknesses as service provision.²⁵²

enabling researchers to trace the intellectual lineage of concepts while using a statistical language that could be parsed by computers and tabulating machines' (Goldenfein and Griffin 2022, No page numbers).

²⁵² There are a great number of critical reviews of citation software systems, for example (Gusenbauer 2019), (Gusenbauer and Haddaway 2020), (Martín-Martín et al. 2021), (Goldenfein and Griffin 2022), (Miller 2019), (Orduna-Malea et al. 2015).

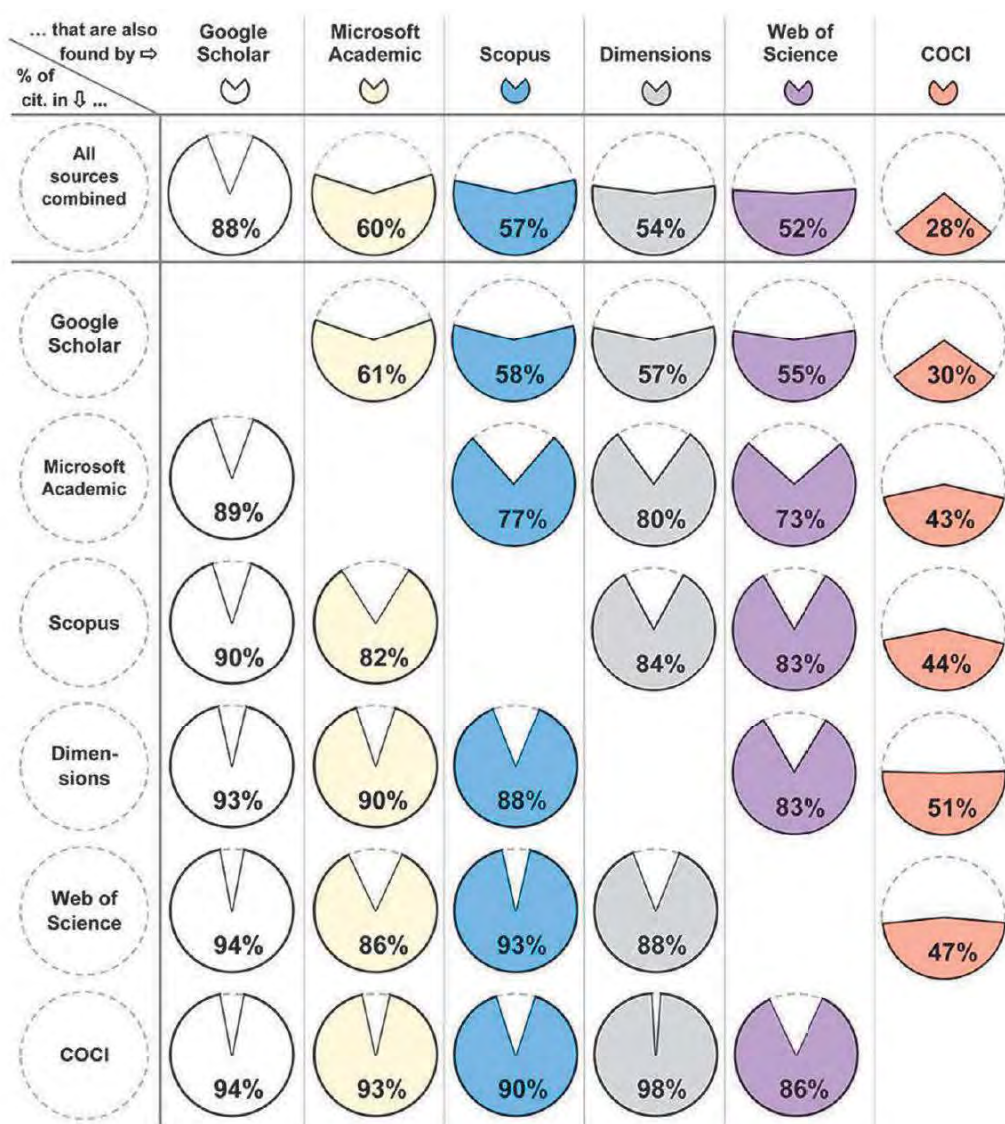


Table 4.2 Percentage of citations found by each database, relative to all citations (first row), and relative to citations found by the other databases (subsequent rows) (Martín-Martín et al. 2021, 882)

4.1.2 Physical primary Records (EBPI): Books, Papers, Records, Objects (BPRO)

Metadata on books and other publications held at libraries follows the long-established discipline of bibliographic control where authorship is bound up with intellectual rights.

Therefore metadata is tightly prescribed and controlled, driven by the intellectual property

concerns of publishers. However, metadata on sources held in archives (BPRO), which carry information about PHL,²⁵³ does not always follow the detail of bibliographic control standards. Nevertheless, metadata in the archive is still universally structured and organised with the author, owner or originator name as the primary key field. In archive work the term 'originator' often has different relevancies in metadata records than author does in library records. In museum work the emphasis is on objects, their placements in time and space and their journeys, and so the creator of the object has relevancies in metadata records similar to but different from those in the archive and the library. These differences between authorship, originator and object creator result in different attitudes towards the sourcing of authorities in each GLAMS sector.

This difference in metadata creation practice in GLAMS sectors contributes to the archivist or curator alone being responsible for the determination and granting of authority to the originating person name when creating BPRO metadata. The archivist or curator is thus free to determine authority by any means within their expertise.²⁵⁴ That cataloguing and ordering systems use 'person name' as a primary field in metadata systems is universal and persistent through time and the reliance on the expertise of the archivist or curator to make determinations is common. The EBP system and NAI-UID system offers an opportunity to

²⁵³ This thesis is not concerned with library holdings, where authorship is established by publishing and where librarians are not called upon to determine authorship, as is the case in archival holdings.

²⁵⁴ 'Archival collections are difficult to manage because by their very nature they are unique, often rare or valuable, frequently fragile, and difficult to decipher. Archivists are able to work with such collections, placing them in historical context and applying their knowledge of archival theory and practice, in addition to preservation of paper and other media to their work. They arrange, describe, and preserve these types of materials every day. Librarians, on the other hand, are used to working mainly with secondary resources, such as books and serials, that can be arranged and described by the call number and cataloging information inherent in the book itself or available through bibliographic databases such as OCLC's WorldCat. Library catalogers can usually download catalog records for their book collections from these bibliographic utilities, and therefore usually have very little original cataloging to do as part of their everyday jobs' (Morris 2009, No page numbers).

recognise the subjective determinations of the archivist and curator by separating those from the instances of person names found in sources.

GLAMS collections all have common features that have dominated their structure and organisation throughout their history. There is also broad commonality in metadata creation with respect to NAR records across GLAMS. These commonalities have allowed GLAMS infrastructures to develop in part by borrowing concepts, ontologies and taxonomies from each other. However, these affect the representation of items in metadata records at libraries and archives differently. In the library, the source of authority for the names of authors in the bibliographic record is that established by the publisher. In the case of a record in the archive, the originator of the record is determined by the archivist using specialist knowledge. Similarly, in the museum the curator determines the authority of the creator(s). So, while NAR files have shared characteristics across GLAMS, the sourcing of the person name itself is different in the library compared to the archive and the museum. This problem grows if related parties are considered (other person names present in sources).

From the perspective of research into PHL, the archivist or curator metadata record is a secondary source of information. The relevant person name information sought by the researcher is the appearance of person names in sources and how this relates to NAR systems (which in the EBP system get their authority from BMD records). Using the EBP system, a match on the names of PHL can be more surely made across GLAMS.

Apart from a relatively few prestigious projects that have achieved the digitisation of whole collections, much more digitisation effort has gone into the digitising of bibliographies, catalogues and indexes of collections and into the development of search engines and

finding aids to explore them, although these remain poorly integrated.²⁵⁵ It is now time to consider not only whether the EBP system extends the utility of GLAMS digital records, but also whether it offers a solution to the integration of digital records across GLAMS.

Archival NAR metadata records commonly reference related persons, sometimes including prosopographical information on them, and the extent of this information is also determined by the specialist archivist or curator. It is not habitually sourced from BMD records, whereas the EBP system would source this information from BMD related Records. In this way genealogical Records can be used to establish reliable familiarly related parties rather than relying on the subjective determinations of specialist archivists or curators. Because the recording in metadata of related persons is the choice of the archivist or curator, they often limit their focus to only significant person names, and therefore disregard names appearing in sources that are thought to be less significant or incidental. EBP recommends collecting every incidence of a person name found in EBPI in Records held at GLAMS, irrespective of the importance or seemed relevance of the person as perceived by the archivist.

The shortcomings of these fundamental aspects in GLAMS Record design and organisational structure in the area of authorities, authorship and related parties were recognised by the Library Linked Data Incubator Group in 2011. It is still the case that considerably more

²⁵⁵ 'Library data today resides in databases which, while they may have Web-facing search interfaces, are not deeply integrated with other data sources on the Web. There is a considerable amount of bibliographic data and other kinds of resources on the Web that share data points such as dates, geographic information, persons, and organizations. In a future Linked Data environment, all these dots could be connected' (Baker et al. 2011, No page numbers).

information is available at archives than that which is findable using current finding and access arrangements.²⁵⁶ EBP seeks to remedy this deficiency.

4.1.3 PHL information (EBPI) and Linked Open Data (LOD)

The person names and prosopographical information that appear throughout and are embedded in the items held in archives form the primary interest of researchers of EBPI.

After the Record itself, the digital representation of EBPI as EBPD is the next concern of EBP and the researcher. The systematisation of LOD throughout all archives is the vehicle through which the concern of the archivist and the interest of the researcher both meet and can be satisfied.²⁵⁷

This thesis takes up the vision of the *Library Linked Data Incubator Group final report* authored by Thomas Baker and twelve other academics in 2011,²⁵⁸ which outlined the

²⁵⁶ 'Some data fields, such as authority-controlled names and subjects, have related records in separate files, and these records have identifiers that could be used to represent those entities in library metadata. However, the data formats in current use do not always support inclusion of these identifiers in records, therefore many of today's library systems do not properly support their use. These identifiers also tend to be managed locally rather than globally, and hence are not expressed as URI.s which would enable linking to them on the Web. The absence of links or insufficient support for them in library systems raises important issues' (Baker et al. 2011, No page numbers).

²⁵⁷ 'A very early step should be the identification of high-priority, low-effort Linked Data projects. By its very nature, Linked Data facilitates an incremental approach to making data available for use on the Web. The data environments of libraries are complex, and attempting to expose that complexity as Linked Data all at once could have limited success. However, some library resources lend themselves to publication as Linked Data without disrupting current systems and services. Among these are authority files (whose members identify things) and controlled term lists. Identification of such "low-hanging fruit" will allow libraries to quickly expand their presence in the Linked Data cloud without changing their workflows elsewhere' (Baker et al. 2011).

²⁵⁸ Thomas Baker, Dublin Core Metadata Initiative, US (W3C Invited Expert); Emmanuelle Bermes, Centre Pompidou, France (W3C Invited Expert); Karen Coyle, Consultant, US (W3C Invited Expert); Gordon Dunsire, Consultant, UK (W3C Invited Expert); Antoine Isaac, Europeana and Vrije Universiteit Amsterdam, Netherlands; Peter Murray, LYRASIS, US (W3C Invited Expert); Michael Panzer, OCLC Online Computer Library Center, Inc., US; Jodi Schneider, DERI Galway at the National University of Ireland, Galway, Ireland; Ross Singer, Talis Group Ltd, UK; Ed Summers, Library of Congress, US; William Waites, University of Edinburgh (School of Informatics), UK; Jeff Young, OCLC Online Computer Library Center, Inc., US; Marcia Zeng, Kent State University, US (W3C Invited Expert) (Baker et al. 2011, No Page numbers).

possibilities for future digital organising and structuring of archival data.²⁵⁹ It was hoped that this would lead to the development of new digital affordances that could make information at archives more digitally accessible, primarily by expanding the use of Uniform Resource Identifiers (URIs).²⁶⁰ This would encourage ‘anyone to contribute unique expertise in a form that can be reused and recombined with the expertise of others’ (Baker et al. 2011, 4).

Essential for the satisfactory development of the 2011 vision of the Library Linked Data Incubator Group is the veracity and appropriateness from a research perspective of the LOD itself.²⁶¹ Archives have made great progress in making available some information on PHL which could be useful in LOD applications through archival authoring systems, but little work has been done on using BMD records to source authorities or in making the

²⁵⁹ ‘Semantic Web technologies conceptualize data in a way that fundamentally differs from the conceptualization underlying the data formats of the twentieth century. Linked Data is primarily about meaning and meaningful relationships between things, while traditional library data formats combine the meaning of data and the structured encoding of data into a single package. The inseparability of meaning from encoding in data formats results in less flexibility for obtaining value from an investment in data. Since the introduction of MARC formats in the 1960s, digital data in libraries has been managed predominantly in the form of “records” that are bounded sets of information stored in files of a precisely specified structure. The Semantic Web and Linked Data, in contrast, structure data as graphs – constructs which, in principle, may be boundless’ (Baker et al. 2011, 11).

²⁶⁰ ‘By using globally unique identifiers to designate works, places, people, events, subjects, and other objects or concepts of interest, libraries will allow resources to be cited across a broad range of data sources and thus make their metadata descriptions more richly accessible. The Internet’s Domain Name System assures stability and trust by putting these identifiers into a regulated and well-understood ownership and maintenance context. This notion is fully compatible with the long-term mandate of libraries. Libraries, and memory institutions generally, are in a unique position to provide trusted metadata for resources of long-term cultural importance as data on the Web’ (Baker et al. 2011, 4).

²⁶¹ ‘The Linked Data approach offers significant advantages over current practices for creating and delivering library data while providing a natural extension to the collaborative sharing models historically employed by libraries. Linked Data and especially Linked Open Data is sharable, extensible, and easily re-usable. It supports multilingual functionality for data and user services, such as the labeling of concepts identified by language-agnostic URIs. These characteristics are inherent in the Linked Data standards and are supported by the use of Web-friendly identifiers for data and concepts. Resources can be described in collaboration with other libraries and linked to data contributed by other communities or even by individuals. Like the linking that takes place today between Web documents, Linked Data allows anyone to contribute unique expertise in a form that can be reused and recombined with the expertise of others. The use of identifiers allows diverse descriptions to refer to the same thing. Through rich linkages with complementary data from trusted sources, libraries can increase the value of their own data beyond the sum of their sources taken individually’ (Baker et al. 2011, 8).

prosopographical information of all PHL contained in archival items available for LOD applications.

4.2 Leading institutions in GLAMS digital development

This section focuses on three leading US institutions in GLAMS that have influenced the development of authority standards both in the US and globally: the Federation of Library Associations and Institutions (IFLA), the International Council on Archives (ICA) and the Expert Group on Archival Description (EGAD). Later affordances (RDA; Describing Archives: A Content Standard, DACS; and RiC-CM) developed out of the detailed groundwork in standards development performed by these organisations. The EBP system will develop best within the sphere of influence of these standards and the organisations which support them.

4.2.1 IFLA Functional Requirements for Bibliographic Records (FRBR)

IFLA²⁶² began developing cataloguing principles in 1961 (called the Paris Principles), quickly followed in 1969 by a commitment to develop international standards in the form and

²⁶² 'The Federation is an independent, international, non-governmental, not-for-profit organization, which advances the interests of library and information associations, libraries and information services, librarians and the communities they serve throughout the world. Formed in 1927, the Federation has its headquarters in The Hague, Netherlands. To achieve its purpose, the Federation seeks to: promote high standards of delivery of library and information services and professional practice, as well as the accessibility, protection, and preservation of documentary cultural heritage. This is done through the enhancement of professional education, the development of professional standards, the dissemination of best practice and the advancement of relevant scientific and professional knowledge; encourage widespread understanding of the value and importance of high quality library and information services in the public, private and voluntary sectors; represent the interests of its Members and library and information organizations and the communities they serve throughout the world.' <https://www.ifla.org/about-us> (Accessed 12 July 2024).

content of bibliographic descriptions. In 2009 IFLA commissioned a report on the modelling of FRBR. The appointed study group brought about the systematic development of FRBR standards that would be adopted by IFLA members. From these foundations follow all subsequent standards in libraries in the US (and by way of exemplar, for other national institutions).²⁶³ The study group reported that it did not have a remit to consider authority records, which it regarded an unfortunate omission, noting that these remained unstandardised in many NAR systems.²⁶⁴ From 2009 FRBR became a major global library standard covering the ‘generic tasks that are performed by users when searching and making use of national bibliographies and library catalogues’ (International Federation of Library Associations and Institutions 2009, 5) (see Table 4.3).

Using the data to find materials that correspond to the user’s stated search criteria (e.g., in the context of a search for all documents on a given subject, or a search for a recording issued under a particular title)

²⁶³ ‘Almost forty years ago the International Federation of Library Associations and Institutions (IFLA) initiated a fundamental re-examination of cataloguing theory and practice on an international level. The first important outcome of that effort was a set of cataloguing principles agreed to at an international conference held in Paris in 1961 that have subsequently come to be known as the Paris Principles. A second key undertaking was initiated at the International Meeting of Cataloguing Experts held in Copenhagen in 1969 with the adoption of a resolution to establish international standards for the form and content of bibliographic descriptions. The first of the standards developed under that resolution, the International Standard Bibliographic Description for Monographic Publications, was published in 1971. In the years that have followed those initial undertakings the Paris Principles and the ISBDs have served as the bibliographic foundation for a variety of new and revised national and international cataloguing codes. (International Federation of Library Associations and Institutions 2009, 1).

²⁶⁴ ‘The model could be extended to cover the additional data that are normally recorded in authority records ... Data associated with persons, corporate bodies, titles, and subjects are analysed only to the extent that they function as headings or index entries for the records describing bibliographic entities. The present study does not analyse those additional data associated with persons, corporate bodies, works, and subjects that are typically recorded only in authority records’ (International Federation of Library Associations and Institutions 2009, 5 & 7).

Using the data retrieved to identify an entity (e.g., to confirm that the document described in a record corresponds to the document sought by the user, or to distinguish between two texts or recordings that have the same title)
Using the data to select an entity that is appropriate to the user's needs (e.g., to select a text in a language the user understands, or to choose a version of a computer program that is compatible with the hardware and operating system available to the user)
Using the data in order to acquire or obtain access to the entity described (e.g., to place a purchase order for a publication, to submit a request for the loan of a copy of a book in a library's collection, or to access online an electronic document stored on a remote computer)

Table 4.3 FRBR founding objectives (International Federation of Library Associations and Institutions 2009, 8)

The FRBR report identifies four entities that give rise to bibliographic records, each of which requires its own NAR entries. These can be different for each entity, and at three points in

the report definitions of appropriate NAR entries for person,²⁶⁵ names of persons²⁶⁶ and multiple persons can be found in the several forms in which a work exists.²⁶⁷ These definition statements direct the structure of metadata elements to ensure conformity and interoperability, both within the FRBR standard and with connected systems.

²⁶⁵ 'The entity defined as person encompasses individuals that are deceased as well as those that are living. Examples:

- p1 Margaret Atwood
- p2 Hans Christian Andersen
- p3 Queen Victoria
- p4 Anatole France

For the purposes of this study persons are treated as entities only to the extent that they are involved in the creation or realization of a work (e.g., as authors, composers, artists, editors, translators, directors, performers, etc.), or are the subject of a work (e.g., as the subject of a biographical or autobiographical work, of a history, etc.). Defining the entity person enables us to name and identify the individual in a consistent manner, independently of how the individual's name appears on or in any particular expression or manifestation of a work. Defining person as an entity also enables us to draw relationships between a specific person and a work or expression of a work for which that person may be responsible, or between a work and the person that is the subject of the work' (International Federation of Library Associations and Institutions 2009, 25).

²⁶⁶ 'The name of a person is the word, character, or group of words and/or characters by which the person is known (e.g., Donald Horne, A. A. Milne, Ellery Queen, etc.). A name may include one or more forenames (or given names), matronymics, patronymics, family names (or surnames), sobriquets, dynastic names, etc. A person may be known by more than one name, or by more than one form of the same name. A bibliographic agency normally selects one of those names as the uniform heading for purposes of consistency in naming and referencing the person. The other names or forms of name may be treated as variant names for the person. In some cases (e.g., in the case of a person who writes under more than one pseudonym, or a person who writes both in an official capacity and as an individual) the bibliographic agency may establish more than one uniform heading for the person' (International Federation of Library Associations and Institutions 2009, 49).

²⁶⁷ 'A work may be created by one or more than one person and/or one or more than one corporate body. Conversely, a person or a corporate body may create one or more than one work. An expression may be realized by one or more than one person and/or corporate body; and a person or corporate body may realize one or more than one expression. A manifestation may be produced by one or more than one person or corporate body; a person or corporate body may produce one or more than one manifestation. An item may be owned by one or more than one person and/or corporate body; a person or corporate body may own one or more than one item' (International Federation of Library Associations and Institutions 2009, 14).

4.2.2 International Council on Archives (ICA)

The ICA²⁶⁸ established four principal archival standards: the General International Standard Archival Description, ISAD(G),²⁶⁹ the International Standard Archival Authority Record for Corporate Bodies, Persons and Families, ISAAR(CPF),²⁷⁰ the International Standard for Describing Functions, ISDF,²⁷¹ and the International Standard for Describing Institutions with Archival Holdings, ISDIAH.²⁷² From these four core standards many others have been developed to extend the reach of standardisation into all areas of the archive. These core standards are now brought together in the Records in Context Conceptual Model (RiC-CM). See Figure 4.1 and Section 4.2.4.

²⁶⁸ 'The ICA Mission statement:

- To encourage and support the development of archives in all countries, in cooperation with other intergovernmental and non-governmental international organisations and businesses. To promote, organise and coordinate the development of best practices and standards and other activities in the field of records, archives and data management.
- To establish, maintain, and strengthen relations between archivists and records/data/information managers in all countries and between all archival and information institutions, professional bodies and other organisations, and with allied professions.
- To support and inspire, worldwide, the work of archival institutions, professional bodies and organisations, public and private, concerned with the administration or preservation of archives, records and data, or with professional training.
- To facilitate the interpretation of records, archives and data by raising their profile and by encouraging their greater use within the established legal frameworks.
- To undertake any relevant activities which support the association's objectives, including but not limited to: organising events, issuing position statements, undertaking programmes, establishing subsidiary bodies, developing training resources, releasing publications, etc.'

<https://www.ica.org/discover-ica/our-mission-our-objectives> (Accessed 23 June 2024).

²⁶⁹ <https://www.ica.org/resource/isadg-general-international-standard-archival-description-second-edition> (Accessed 23 June 2024).

²⁷⁰ <https://www.ica.org/resource/isaar-cpf-international-standard-archival-authority-record-for-corporate-bodies-persons-and-families-2nd-edition> (Accessed 23 June 2024).

²⁷¹ <https://www.ica.org/resource/isdf-international-standard-for-describing-functions> (Accessed 23 June 2024).

²⁷² <https://www.ica.org/resource/isdiah-international-standard-for-describing-institutions-with-archival-holdings> (Accessed 23 June 2024).

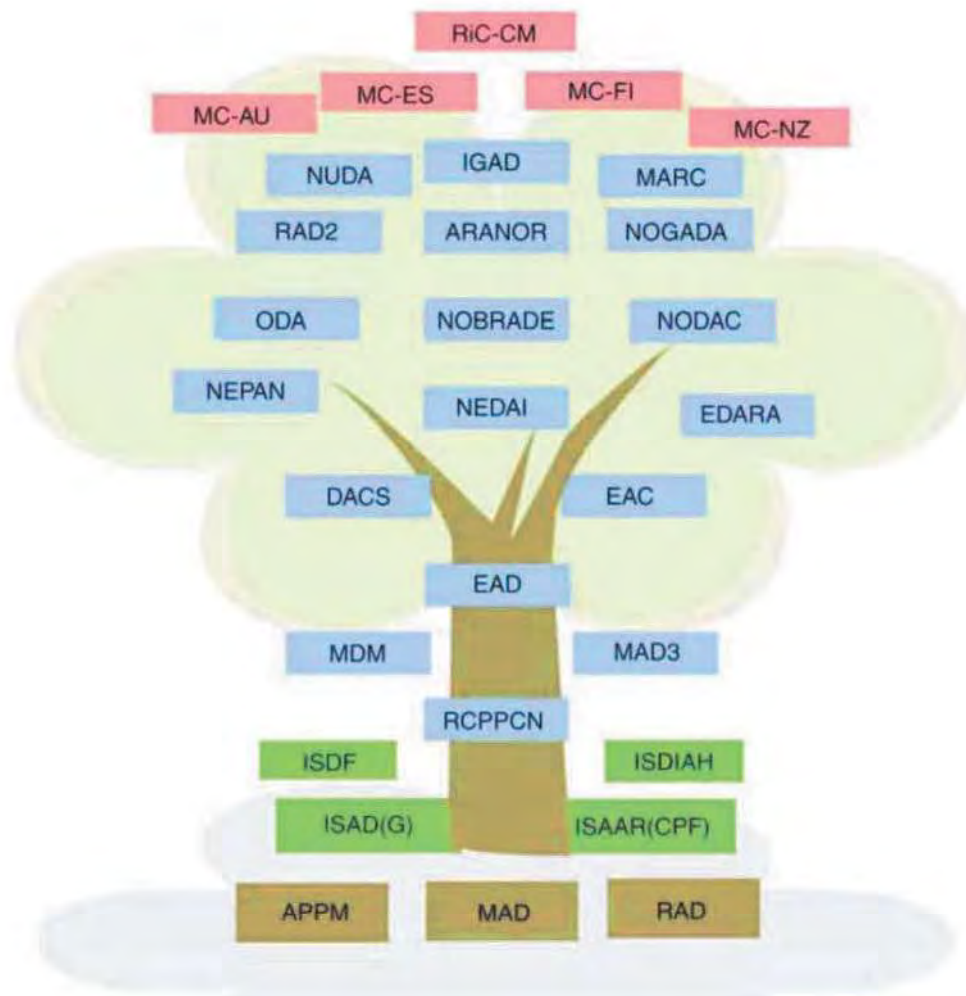


Figure 4.1 Graphic representation of the history of the development of norms and models of archival description since the end of the 1980s (Llanes-Padrón and Pastor-Sánchez 2017, 391)

The ICA found that, due to regional and cultural variances, it was not possible to establish rigid global standards. Instead, the ICA international standard allows for each user to determine how best to accommodate often essential local variations into the standard:

This standard provides general guidance for the preparation of archival descriptions. It is to be used in conjunction with existing national standards regulations or as the basis for the development of other national standards. (International Council on Archives 1999, 7)

Oversight of the development of standards in the ICA is undertaken by its Expert Group on Archival Description (EGAD).²⁷³ The ICA conducted a comprehensive survey in 2021 which showed that the recognition of the importance of standards is universal (Figure 4.2).²⁷⁴

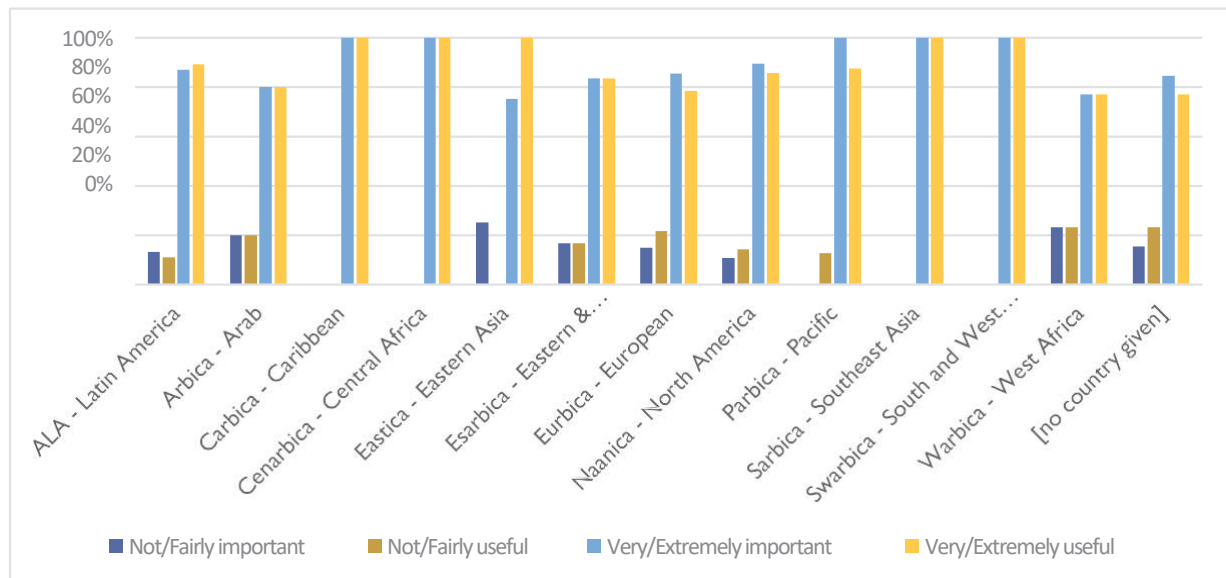


Figure 4.2 How important and useful are standards? (International Council on Archives 2021, 5)

However, the take-up of the four principal standards (prevailing in 2021) is less universal, varying between the four related but different standards, which taken together provide a standard framework for the entire archive and its collections. This is shown in Figure 4.3.

²⁷³ 'Developing international consensus on a standard for archival description is a daunting challenge. Cultural differences coupled with differing theories and practices are at the core of this challenge. The members of the EGAD represent many (though certainly not all) of these differences. At the same time, they share a common commitment to developing a shared standard that respects and accommodates the past practices, and that respects and accommodates differences while remaining intellectually coherent and workable. EGAD also recognizes that developing a consensus will necessarily be an ongoing process, a field of negotiation' (International Council on Archives 2023, 11-12).

²⁷⁴ 'Archivists have been trained to alter MARC records to fit their collections, to conduct historical research to learn more about the collection creators and relevant time period or events covered in the materials themselves, and to arrange the collections and create finding aids, or guides, to those collections so that they are easier to use' (Morris 2009, No page numbers).

The general standard applicable to documents (items) held at archives is ISAD(G).²⁷⁵

ISAAR(CPF), 2nd edition, is the standard for authority records. This standard has direct implications for EBP and the NAI-UID system and therefore it will be considered in detail below. Finally there is ISDIAH. It can be seen that progress in securing archives is much higher than when implementing international standards in collected items.



Figure 4.3 How far are these standards implemented in your archive? (International Council on Archives 2021, 6)

The reason for relatively low implementation of ISAAR(CPF) is probably because of the presence of hard-to-replace legacy national standards (Figure 4.4).

²⁷⁵ 'International Standard Archival Description (General) or ISAD (G) is an international standard which provides guidelines for creating the content of an archival description. It promotes the creation of consistent and appropriate descriptions, aiding the retrieval and exchange of information, and the integration of descriptions into a unified information system. ISAD (G) sets out a list of elements which are considered necessary for an archival description, and rules that should be followed when writing a description. ISAD (G) identifies and describes what kind of information should be included in an archival description and whether this description is in written, printed, or electronic form.' <https://archiveshub.jisc.ac.uk/isadg> (Accessed 12 July 2024).

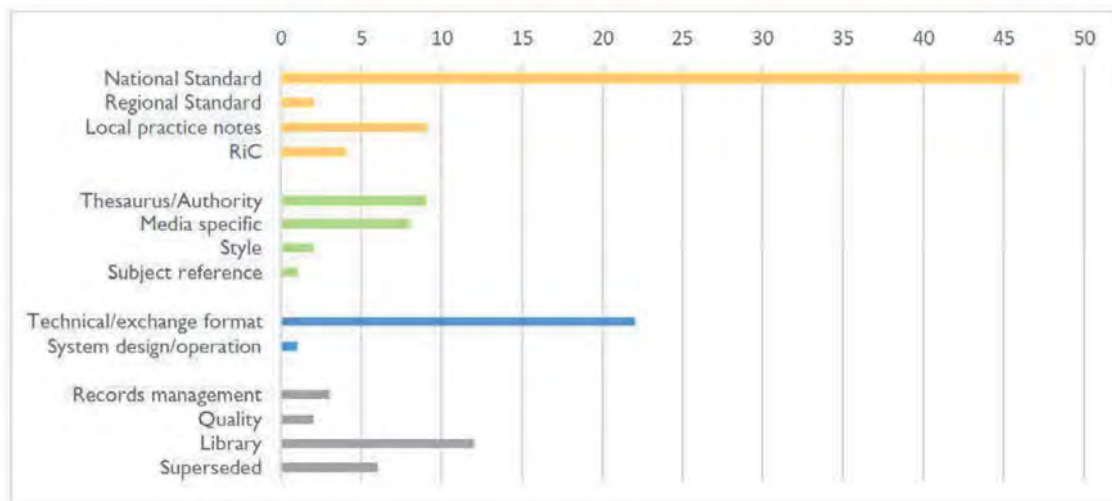


Figure 4.4 What other standards do you use to describe your archives? (International Council on Archives 2021, 6)

The ICA has been very proactive in the development of standards in archive work, but take-up is more challenging than adoption. Recent consolidation of standards into the RiC standard will go some way to address the difficulties in take-up as far as these relate to relieving the burden of regulation on what is a mixed community. Different regions are able to bring greater or lesser resources to bear on regulation of provision, and, paradoxically, early digitisers face the sometimes daunting task of now having to adapt or replace prior digitising efforts, which can be challenging.

4.3 Metadata and data control in GLAMS

There are many metadata schemas in popular use.²⁷⁶ The formal and disciplined structuring of metadata is necessary for the higher elements of the digital world to function effectively

²⁷⁶ 'The Digital Curation Centre (DCC) is a world-leading centre of expertise in digital information curation with a focus on building capacity, capability and skills for research data management.' <https://www.dcc.ac.uk/guidance/standards/metadata/list> (Accessed 27 July 2024), where there is also a comprehensive list of standards relating to curation.

and efficiently, such as XML, RDF, RDFS and OWL schemas,²⁷⁷ and for web operations such as SKOS.²⁷⁸ There are six types of inter-related metadata in the GLAMS area (see Table 4.4) and the ‘descriptive metadata’ type offers scope, via extensibility and using MARC21 markup language,²⁷⁹ for metadata schemas to include EBP and use the NAI-UID indexing system.

²⁷⁷ ‘[W]e examined the modeling aspects of the Semantic Web: How do you represent information in such a way that it is responsive to a web environment? The basic principles underlying the Semantic Web—the AAA slogan, the Nonunique Naming assumption, and the Open World assumption—are constraints placed on a representation system if it wants to function as the foundation of a World Wide Web of information. These constraints have led to the main design decisions for the Semantic Web languages of RDF, RDFS, and OWL’ (Allemang and Hendler 2011a, 337).

²⁷⁸ ‘The applications we discussed in this book demonstrate how a modest amount of information, represented flexibly so that it can be merged in novel ways, provides a new dynamic for information distribution and sharing. SKOS allows thesaurus managers around the globe to share, connect, and compare terminology. QUDT aligns multiple applications so that their measurable quantities can be combined and compared. OBO Ontologies coordinate efforts of independent life sciences researchers around the globe’ (Allemang and Hendler 2011a, 337).

²⁷⁹ ‘The acronym MARC stands for Machine-Readable Cataloging, and “MARC 21” is the name of the current publication of those standards. MARC was originally developed in the 1960s for the recording of bibliographic records in machine-readable form, but several other formats developed, including an authorities format’ (Maxwell 2001, 2.11). See also (Maxwell 2001, 12) and the MARC21 Authority website, <https://www.loc.gov/marc/authority/adintro.html> (Accessed 28 July 2024).

Metadata Type	Example Properties	Primary Uses
Descriptive metadata	Title Author Subject Genre Publication date	Discovery Display Interoperability
Technical metadata	File type File size Creation date/time Compression scheme	Interoperability Digital object management Preservation
Preservation metadata	Checksum Preservation event	Interoperability Digital object management Preservation
Rights metadata	Copyright status License terms Rights holder	Interoperability Digital object management
Structural metadata	Sequence Place in hierarchy	Navigation
Markup languages	Paragraph Heading List Name Date	Navigation Interoperability

Table 4.4 Metadata types (Riley 2017, 7)²⁸⁰

²⁸⁰ See also (Riley 2017, 6-7): ‘These various categories of metadata support different use cases in information systems. Discovery is perhaps the most common, with structured metadata allowing users to search for or browse to find resources or information of interest. Many metadata properties are useful to display to users to aid in identification or understanding of a resource. Interoperability, the effective exchange of content between systems, relies on metadata describing that content so that the systems involved can effectively profile incoming material and match it to their internal structures. Metadata supports digital-object management by providing the information needed to render digital content appropriately or deliver the appropriate version to match a user need. Preservation is achieved through creating metadata that allows the verification of the integrity of content after transfer and at other notable points, and signaling when preservation actions such as a format migration or an integrity check should be undertaken. Finally, metadata supports navigation within parts of items, for example, from one page or section to the next, and among different versions of objects, such as varying resolutions of photographic images.’

Metadata schemas in popular use include Dublin Core, which is a MARC (Machine Readable Cataloguing) descriptive archival metadata schema, EAD (Encoded Archival Description),²⁸¹ RDA, and MARC's eventual replacement, BIBFRAME.²⁸² Descriptive metadata is at the heart of the archival cataloguing system, and it is the point of interaction with the recommended Archival Authority Index component of the NAI-UID system. It is therefore important to explore the extent to which existing archival metadata affordances can accommodate EBPD and the NAI-UID system, if it were adopted.

Mirna Willer and Gordon Dunsire explain the importance and centrality of archival metadata to the digital world and its ability to point to sources.²⁸³ They also set out unambiguously the potential for URIs to replace or enhance the local archival authority record identifier in metadata files.²⁸⁴ There are problems in using archival metadata vocabularies in the LOD environment, as the W3C Cluster Archives report explains (Baker et al. 2011, 7). It is now

²⁸¹ 'Encoded Archival Description (EAD) is the international metadata transmission standard for hierarchical descriptions of archival records. Developed by the EAD Working Group of the Society of American Archivists and first published in 1998, EAD is an Extensible Markup Language (XML) format used by archivists around the globe' (Society of American Archivists 2023, 8).

²⁸² 'BIBFRAME is short for Bibliographic Framework. It began as an LC initiative in 2011 to transition from a legacy, MARC-based environment to one that fully integrates with and reaps the benefits of the World Wide Web. BIBFRAME is the foundation for the future of bibliographic description; it will become the primary means of bibliographic data exchange; and it will replace the MARC Format. BIBFRAME's primary benefit to the community of knowledge seekers is its ability to enhance information exploration through the use of links and World Wide Web technologies, creating a virtual "stack browsing" experience while improving on physical browsing' (Library of Congress 2019, 6).

²⁸³ 'What we do know is that the existence of many types of metadata will prove critical to the continued online and intellectual accessibility and utility of digital resources and the information objects that they contain, as well as the original objects and collections to which they relate. In this sense, metadata provides us with the Rosetta stone that will make it possible to decode information objects and their transformation into knowledge in the cultural heritage information systems of the future.' (Gilliland 2008, 19, quoted in (Willer and Dunsire 2013, 65).

²⁸⁴ 'Value vocabularies are defined as controlled lists of allowed values for an element; that means that value vocabularies define resources that are used as values for specific elements in a metadata record. For example, an author element in a metadata record may contain the author's name as a literal string, or the identifier (ID) of the record for that author in a name authority file. In the linked data environment each entry in such an authority file would be represented by its own URI, and instead of the record ID the value in the metadata element would be that assigned URL' (Willer and Dunsire 2013, 272).

timely to introduce the concepts of EBP and the NAI-UID system into standards development in GLAMS, while systematisation of AAC is still in progress (see Table 4.5).

Missing vocabularies
Sometimes specific (physical state of original in a preservation context), sometimes general (need vocabularies for preservation data), but no vocabulary for the function of data elements
Data incompatibilities or lack
Current data is free text, but contains quantitative information that needs to be pulled out
Data needs to be qualified as 'estimated' or 'derived', so users know it is not precise (this is possibly a vocabulary issue)
Current practice does not include rich relationships, just 'related', so there is no source of relationships
No examples in our community domain that we can follow
Lack of information on how to create a data model
No community guidance on which technologies and vocabularies to use
Is linked data scalable to the size we need?
Is linked data appropriate for highly hierarchical data models?
No systems available on market for linked data creation and use
Open source solutions available are in an unfinished state

Table 4.5 Problems in using archival metadata vocabularies in the LOD environment Karen Coyle and Emmanuelle Bernes, W3C, Cluster Archives, September 2011

([https://www.w3.org/2005/Incubator/ldl/wikil/Cluster Archives](https://www.w3.org/2005/Incubator/ldl/wikil/Cluster%20Archives), quoted in (Willer and Dunsire 2013, 267)

International and national standards in the library have been in place since the early 1970s under the IFLA umbrella. At the core of the many GLAMS standards that have been developed, rolled out, then later replaced by improved and more integrated applications is the metadata Machine Readable Cataloguing format (MARC),²⁸⁵ and at the time of writing the latest version is MARC21.²⁸⁶ It also underpins the future affordance, BIBFRAME.²⁸⁷

Resource Description and Access (RDA)²⁸⁸ is a new concept in international metadata standard affordance because it is an online dynamic standard (offered exclusively through the online 'RDAToolkit') and aimed to be of use across GLAMS²⁸⁹ in linked data

²⁸⁵ 'The MARC21 bibliographic format is used for describing the items that libraries hold. It is made up of several hundred fields, though a much smaller core set is used most frequently. These include fields for various types of titles, authorship of works by people or groups, edition and publication information, physical description, series, notes, and subject and genre terms. The MARC21 Authority format is used for documenting controlled terms for people, corporate bodies, work titles, subjects, and genres. These controlled terms are then used as entries in the appropriate fields in MARC21 bibliographic records, to provide consistency in these records and aid discovery. The MARC Authority format includes fields for encoding the controlled heading, which often include additional metadata about the entity—occupation for a person, for example; address for a person or corporate body; or key for a musical work. It further includes fields for encoding alternate forms of a name for the entity and making notes that document why the particular form of the controlled heading was chosen' (Riley 2017, 28). See also (Library of Congress 2009).

²⁸⁶ 'MARC is actually a family of formats, with different implementations of ISO 2709 used in different countries. Most notable are the MARC21 formats maintained by the Library of Congress, which are in use in the United States, Canada, and much of the English-speaking world. MARC21 is composed of five formats: MARC21 Bibliographic, MARC21 Authority, MARC21 Holdings, MARC21 Classification, and MARC21 Community Information. The primary schema for encoding bibliographic and authority records since the late 1960s and early 1970s has been the MARC: Machine Readable Cataloguing format. Although its versions have proliferated in various forms of national implementation, they all share the same structure specified by ISO 2709, an international standard which has enabled and facilitated the widespread exchange and reuse of bibliographic data we enjoy today. The standard was developed from the record structure of LC MARC which was standardized as ANSI Z39.2 in 1971' (Willer and Dunsire 2013, 51). See also (Riley 2017, 27).

²⁸⁷ 'BIBFRAME provides a foundation for the future of bibliographic description that is grounded in Linked Data techniques' (Library of Congress 2019, 4).

²⁸⁸ 'RDA is a key step in the improvement of resource discovery because it guides the recording of data. The production of well-formed data is a vital piece of the infrastructure to support search and retrieval. RDA data alone will not improve navigation and display because the data must be used appropriately by well-designed applications, search engines and interfaces. Nevertheless, the recording of clear, unambiguous, well-structured data is an essential step in the improvement of resource discovery for the user' (Oliver 2021, 11).

²⁸⁹ 'RDA was designed to make bibliographical information usable as data. It was not designed for one particular encoding scheme; the intention is that RDA data should be suitable for use with a range of different encoding schemes. RDA is intended to be the basis for a metadata element set that will make data visible and usable in library catalogs, on the World Wide Web or in a Semantic Web environment' (Oliver 2021, 5).

environments. It was first published in 2010²⁹⁰ and was designed to bring together all FRBR metadata standards relating to IFLA relevant conceptual models in force at the time of its launch.²⁹¹ ICA, IFLA and the International Council on Museums (ICOM)²⁹² are driving the next generation of metadata development in GLAMS.

4.3.1 Universal Bibliographic Control (UBC)

Understanding the development of digitisation in the library begins with the concept of UBC.²⁹³ The work of Willer and Dunsire has informed this section; Dunsire was a member of the 2011 Library Linked Data Incubator Group. A defining feature of the development of digital UBC is that integration into one global system is neither desirable nor practical, at either the metadata or the standards levels. However, because national systems and standards are common in the field of bibliographical control, linking and pathing between

²⁹⁰ 'Resource Description and Access is an international metadata standard designed to enable the discovery of library and cultural heritage resources in both traditional and linked data environments. It evolved out of the Anglo-American Cataloguing Rules, 2nd edition (AACR2), but RDA is quite different. It presents a new way of thinking about bibliographic data. It is based on a theoretical framework, it is designed as a standard for the digital environment, and it is developed as a global standard appropriate for use in many contexts ... For the cataloging community, the publication of RDA in 2010 marked a new approach to the recording of bibliographic data but it also introduced a new way of using the standard. RDA was designed to be used as an online tool. The content of the standard was published as part of an online web-based tool, RDA Toolkit. The text of RDA had been prepared as a series of documents and these were then transferred into specially designed software in 2010' (Oliver 2021, 1.1).

²⁹¹ 'In 2010, when RDA was first published, it was aligned with the first two of IFLA's conceptual models, Functional Requirements for Bibliographic Records (FRBR) and Functional Requirements for Authority Data (FRAD). In 2015, some additions were made to RDA so that it also aligned with Functional Requirements for Subject Authority Data (FRSAD). FRAD and FRSAD were extensions of the FRBR model. Thus, as of 2015, RDA was essentially aligned with all three IFLA models that were in force at that time' (Oliver 2021, 3).

²⁹² <https://icom.museum/en/resources/standards-guidelines> (Accessed 29 July 2024).

²⁹³ 'The concept of Universal Bibliographic Control is based on the objective of promotion of a world-wide system for control and exchange of bibliographic information. The purpose of the system is to make universally and promptly available, in a form which is internationally acceptable, basic bibliographic data on all publications in all countries' (Willer and Dunsire 2013, 3).

national systems have become commonplace, enabling global frameworks to emerge.²⁹⁴

Another key feature in the early development phase of UBC was the adoption of Conceptual Reference Models (CRM) that set out the objectives, definitions and parameters for metadata system design and adoption. CRMs enable the history of the development of library system digitisation to be studied and analysed in great detail.²⁹⁵ The UBC CRMs were early adopters of relational database methodologies, with the users of library information as the focus.²⁹⁶

4.3.2 Archival Authority Control (AAC)

The first difficulty in AAC is the complex nature of archive texts themselves. Texts have varied lives of equally varied complexities, and all of the many possible manifestations of a text can give rise to separate but related copies in other archives, often leading to an abundance of name variances recorded at each manifestation.

²⁹⁴ 'At the international level the integration of national bibliographic agencies to form the total system depends upon universal recognition and acceptance that each national bibliographic agency is the organization responsible for creating the contents: the elements to be included in the authoritative bibliographic record of the publications of its own country. In other words, UBC has been based on a system by which the comprehensive bibliographic record of a publication is made once in a country of its origin, in accordance with the international standards which are applicable in both manual and in mechanized systems; and is then available speedily, in a physical form which is also internationally acceptable' (Willer and Dunsire 2013, 5).

²⁹⁵ 'The conceptual models for bibliographic and authority data in particular have made an immense impact on the "thinking" behind the bibliographic universe and its principles and standards, catalogue production and services, and professional and user education and guidance, as well as on concepts of their "universe" in other communities, from museums, archives and publishers to wider Internet groups' (Willer and Dunsire 2013, 14).

²⁹⁶ 'The conceptual models were developed from two strategic decisions. The first one was the choice of methodology for modelling the "population" of the bibliographic universe. The methodology was adopted from relational database systems where data analysis techniques require the key objects or entities that have particular functions within the "universe" in question to be defined. The second decision was to agree about what those functions would be: they would be based on user requirements. Thus, each entity was defined in the context of the user as "the key object of interest to users of bibliographic data". At the foundation of the modelling was a re-examination of the relationship between individual data elements in the record and the needs of the user' (Willer and Dunsire 2013, 14).

Frequently texts are grouped in the archive into ‘families’ called ‘fonds’²⁹⁷ (Figure 4.5).

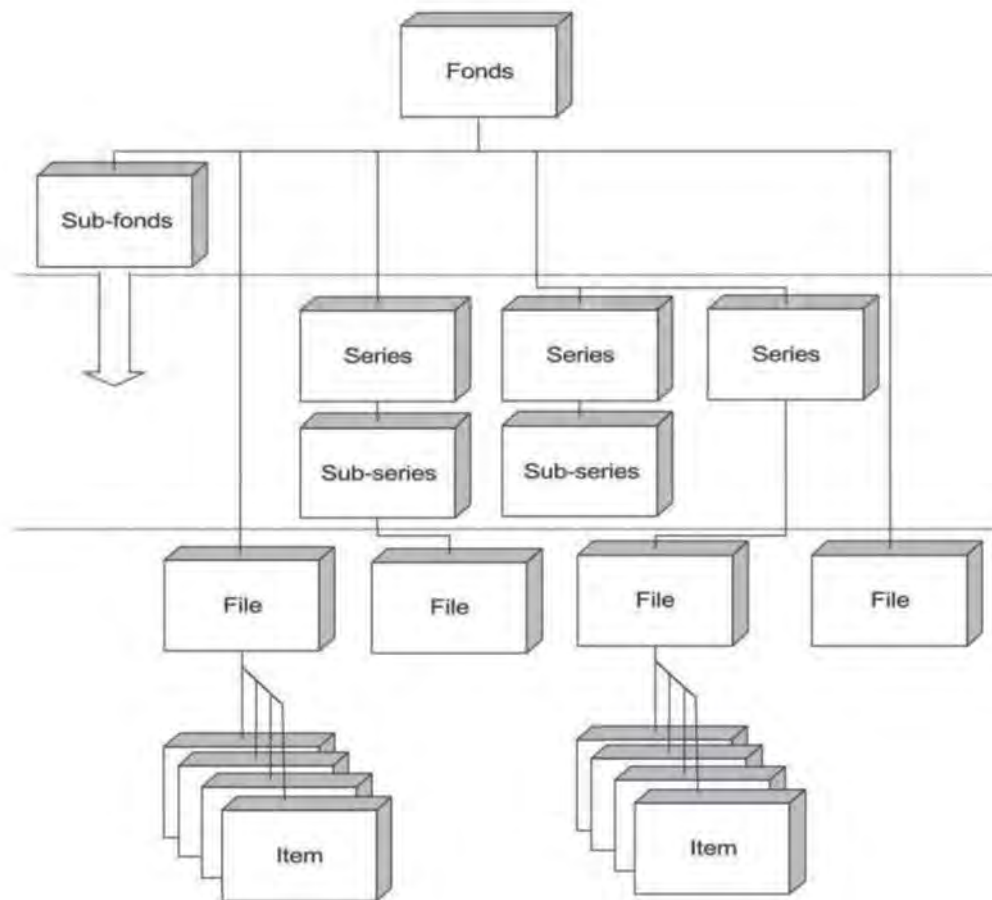


Figure 4.5 Model of the levels of arrangement of a fonds (International Council on Archives 1999, 37)

The presence in the archive of fonds brings with it complexity, and that complexity is magnified when authority records are added to the fonds system (Figure 4.6).

²⁹⁷ ‘Respect des fonds’ is now open to challenge in a digital context. See (Millar 2002).

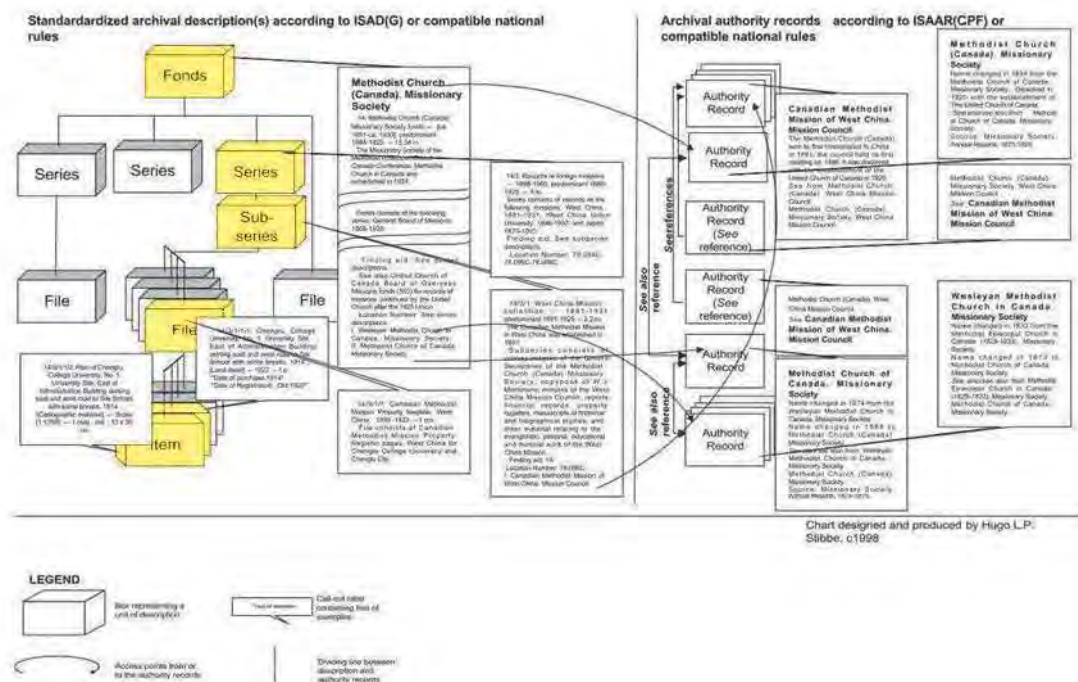


Figure 4.6 Relationship between descriptive and authoritative records (International Council on Archives 1999, 38)

Further complexities easily arise, such as variability in the appearance of person names in multiple editions, and the complications of multiple authorship (for example, later revisions or corrections).²⁹⁸

The fonds organising principle in the AAC is that data is organised hierarchically and rigidly around the concept of the originator. This presents difficulties for the digitising archivist in best practice in the use of standards when producing metadata. That archivists successfully manage, and researchers (if with some difficulty) negotiate, these complexities gives encouragement that the simplicity and Record focused approach of the EBP and the NAI-UID

²⁹⁸ 'Once in the bibliographical universe, a text can become an ancestor of a huge family of other texts related to it in an immense variety of ways and degrees; and any of these near or remote relatives, versions and derivatives, might have one or many published appearances, and start its own family of related texts' (Wilson 1983, 4-5).

indexing system can make a positive contribution by helping to relate persons in the fonds system to persons in the NAI-UID system.

Metadata of bibliographical records is already rich in unattributed prosopographical data (because it relies on the affirmation of the archivist). If the NAI-UID system were to be adopted in GLAMS,²⁹⁹ because it is based on BMD records in a national authority system AAC metadata would be enhanced, without disturbing the fonds system.

4.4 Standards and conceptual models in GLAMS

Carlo Bianchini, writing 2022, points out that the approach to cataloguing has traditionally been based on object characteristics that ‘fit’ the natural structure of the catalogue with, over time, greater refinements.³⁰⁰ He goes on to indicate a vision for the future where data professionals in GLAMS embrace LOD and the Semantic Web to provide more complex modelling of the archive.³⁰¹

²⁹⁹ ‘It is not the role of archivists and similar data stewards to limit which queries can be made beyond the legal restrictions (such as embargos), but instead to facilitate discovery and reuse of resources through adequate and complete enough archival descriptions that maximize the relationships established by resources, enabling users to construct their queries in a manner that is not pre-determined. This notion reinforces the importance of developing a model for authenticity in archival description that can serve as a guideline for these professionals’ (Pacheco, Da Silva, and De Freitas 2023, 637).

³⁰⁰ ‘The traditional approach to cataloguing has always been based on the investigation of the characteristics of the objects in collections that are most relevant to the construction of catalogues. The evolution of cataloguing theory has therefore led to the identification and analysis of entities with those characteristics and to the construction of a logical model capable of explaining bibliographic phenomena in an increasingly refined manner. That approach also led to the definition of RiC-CM and CIDOC-CRM models’ (Bianchini 2022, Abstract). He could have added ISAAC (CPF) and DACS.

³⁰¹ ‘In addition to this valid point of view, a second perspective is proposed, which takes into account the entities identified in the logical models developed in the library, archive and museum models as part of the much larger, richer and more numerous ontologies of the semantic web, represented by the Linked Open Data Cloud. In this perspective, the logical models of libraries, archives, and museums can be seen as some of the possible infinite modelling of web entities, constructed in the light of the principles and tradition of each subject area. This new perspective helps to better clarify the role of data professionals, the concept of metadata, the characteristics of logical models and to take a unified view of the bibliographic, archival and museum universes’ (Bianchini 2022, Abstract).

The section begins with a look back at the work of Name Authority Cooperative Project (NACO) in libraries, which began in 1977, because it provides the bedrock for the rest of the section, which concentrates on those aspects of recent developments that support NAR standardisation. From 2003 onwards, ISAAR(CPF) was the most common universal standard in libraries that embeds NAR principles and recommendations. In 2012 BIBFRAME emerged. In 2013 DACS learnt from ISAAR(CPF) and consolidated several previous archival standards. In 2016 the ICA then introduced RiC-CM. In museums, in 1998 CIDOC was formed and in 2021 it launched its new CIDOC-CRM. Efforts are now being made to harmonise all of these standards in cultural heritage sources.³⁰² The future of humanities is most likely to be characterised by more consolidation of standards and their affordances, with a corresponding natural flexibility to enable cultural and regional differences to be fully recognised. If this consolidation does go ahead in an expanding digital universe, then it would be appropriate to include consideration of the EBP and NAI-UID concepts in that development.

³⁰² 'In a world where data silos have been broken for years, conceptual models and ontologies are used for information integration, and promoting interoperability between data of related domains. In the cultural heritage field, CIDOC Conceptual Reference Model (CIDOC-CRM) and Europeana Data Model (EDM) are widely used for integrating heterogeneous data and enabling effective data sharing. In the context of archives, recently, a similar conceptual model, called the Records in Contexts Conceptual Model (RiC-CM), was proposed for this purpose. RiC-CM is a new, high-level conceptual model for the definition of archives, records and other related entities, which is gradually implemented by archival organizations, along with the corresponding ontology (RiC-O). CIDOC-CRM is widely acknowledged as a suitable framework for modeling the interconnectedness and mappings between diverse cultural heritage sources. Works in this direction include VRA and Dublin Core mappings to CIDOC-CRM. The ICA also acknowledges the necessity for mappings between certain entities or properties of RiC-O and other models such as CIDOC-CRM, IFLA-LRM, PREMIS, etc. This highlights the importance of establishing connections and relationships between different cultural heritage data models for enhanced interoperability' (Bountouri et al. 2023, 90-91).

4.4.1 Name Authority Records (NAR)

The leading organisation in NAR work in the US is the Name Authority Cooperative Project (NACO), established in 1977.³⁰³ NACO maintains the LC/NACO Authority File³⁰⁴ of over two million authority records.³⁰⁵ The working assumption made by the IFLA and its related standards bodies with regard to the NAR record making of archivists³⁰⁶ is that it should establish authenticity in name records.³⁰⁷ It should also provide a centralised and comprehensive finding aid operating across multiple institutions.³⁰⁸

NACO argues that the specialist archivist, in their determination of authority, gives ‘the proof of authenticity needed’.³⁰⁹ EBP counters this definition, instead asserting that an independent National Authority Record based on BMD records and embracing all PHL (and

³⁰³ ‘In the United States, libraries have heavily relied on the Library of Congress/NACO Name Authority File (LCNAF) (<http://id.loc.gov/authorities/names>) to provide them with the necessary controlled access points and cross-references to fulfill that need. Given the never-ending amount of materials that libraries acquire, that is a tall order, particularly with the evolution of digitized and Born Digital materials and their newly adopted and developed repositories’ (Lampron and Wacker 2019).

³⁰⁴ ‘LC/NACO Authority File (often informally called the LCNAF, the NAF, or the NACO File) is a cooperatively-maintained authority file of over 10 million descriptions for persons, families, corporate bodies, jurisdictions, works, and expressions, expressed using the MARC Authority Format’ (Cannan, Frank, and Hawkins 2019, 39).

³⁰⁵ ‘[NACO] now encompasses some 395 institutions that have collectively developed and maintained a database of more than 2,000,000 authority records in addition to the more than 3,500,000 records created by Library of Congress staff ... with a membership now including institutions from all but four of the 50 states comprising the U.S., and including 43 institutions in 16 countries within Europe, Africa, Oceania, Asia, and Latin America’ (Byrum 2004, 238)

³⁰⁶ ‘The Anglo-American Cataloguing Rules (AACR2) recognize three types of names. These are personal names (AACR2 chapter 22), corporate names (AACR2 chapter 24), and geographic names (AACR2 chapter 23). All of these names are subject to authority control because, as access points, it is desirable that they take one form and one form only so that users may have the expectation of finding everything associated with a name by entering only one search’ (Maxwell 2001, 71).

³⁰⁷ ‘[S]uch a record is made by an expert cataloguer and bibliographer following the demands for uniformity provided by the use of international standards, gives it the proof of authenticity needed for it to be reused in local catalogues and information services’ (Willer and Dunsire 2013, 4).

³⁰⁸ ‘Ideally the user, knowing the original form of the name of an author, should find all their works brought together or collocated under that name in any of the catalogues or bibliographies they would be consulting’ (Willer and Dunsire 2013, 6).

³⁰⁹ ‘[A] full authority record contains notes justifying the choice and form of the heading. Just as a good article or book will cite its sources of information, so a cataloguer will cite the source of information both for the form of the name (which may come from the title page of a book) and for other information, such as dates of birth and death (which might come from a reference source)’ (Maxwell 2001, 5-6).

not just those names of interest to archivists) should be adopted and used by archivists when creating archival records in respect of persons, corporate bodies and families.³¹⁰

An IFLA/ICA working party, when considering revisions to ISAAR(CPF) in 2003, ran into difficulty when considering international standard specifications that would allow the standardisation and sharing of archival authority data, and as a consequence the subject was dropped and has not been taken up again since.³¹¹ New ways of building authority structures are today emerging³¹² and there is a new interest in incorporating others (including researchers) in the enterprise of building up reliable information relating to archival metadata.³¹³ The EBP system recognises national primacy in person name authority indexing and the system affords a structured and universally reliable solution to the many problems of person name authentication in GLAMS. Importantly, the EBP system supports

³¹⁰ 'The elements of the entity-relationship models – that is, entities, their attributes and relationships – were derived from data typically found in bibliographic and authority records as specified by IFLA standards, guidelines and formats (specifically UNIMARC)' (Willer and Dunsire 2013, 15).

³¹¹ 'The problems in sharing authority data, however, seemed to outnumber the possible benefits, as they referred to the different treatment of, for example, the change of the name of a corporate body throughout the course of its existence because of the need for context, and because of all the documents it produced, which therefore have to be referred to in finding aids or catalogues. This whole chapter was dropped in the following versions, while the contacts between the two communities [the IFLA and the ICA] on the topic were not officially continued' (Willer and Dunsire 2013, 265-266).

³¹² 'Another early RDF vocabulary on the open Web is Friend of a Friend (FOAF). This vocabulary provides for descriptive metadata on people and organizations, along with their attributes and relationships. FOAF classes include Person, Organization, Group, and Project, which make use of properties such as name, title, and member. The FOAF Core defines a small number of primary classes and properties designed to be a baseline for further refinement and specification. FOAF also includes a list of classes and properties related to how people and organizations interact with the social Web, including terms such as PersonalProfileDocument and accountName. FOAF is primarily used in systems to identify people and organizations and provide basic information about them. The vocabulary does support some more advanced features, however, such as denoting the individuals who are members of a group or documenting personal interests' (Riley 2017, 25).

³¹³ '[T]he emerging culture of openness and interconnectedness of metadata is leading to a redefined definition of "authoritative" or "good" metadata. The Web has opened up new opportunities for previously marginalized voices to share their knowledge. Informed enthusiast communities exist online for nearly every topic, and these individuals can frequently provide far better metadata than organizations that are tasked with managing content but that lack this subject expertise. This process is sometimes known as "crowdsourcing" or "trusting the wisdom of the crowd." Intelligent systems can combine this user-generated metadata with metadata from more traditional sources in a way that is sensible to users. To facilitate this deeper integration of metadata, we can likely expect interfaces of systems designed for laypeople to become incredibly streamlined and user friendly, building on activities this community is already motivated to do' (Riley 2017, 40).

and provides additional assurance to the unstructured, ‘the archivist knows best’, current approach to NAR in GLAMS.³¹⁴

The Virtual International Authority File (VIAF)³¹⁵ of OCLC³¹⁶ (Online Computer Library Center, the owners of WorldCat) links together into one NAR service the name authority files of over forty archives from over thirty countries. OCLC has an extensive ongoing project which started in 2012 linking GLAMS MARC metadata access points using LOD technologies,³¹⁷ and it is now the accepted international organisation for the allocation of URIs³¹⁸ to name clusters.³¹⁹ This work is not without its challenges, many of which derive from the inherent nature of person name variabilities noted throughout this thesis.³²⁰ In 2021, Tang (Cindy)

³¹⁴ ‘There is usually room for cataloguer judgment in the choice of form for a given name or subject, so different catalogers might arrive at differing headings for the same name’ (Maxwell 2001, 1.3).

³¹⁵ ‘Exchanging information and data requires standards, at both the national and international level, for description, identification, and data format. Nowadays, a pillar of UBC is VIAF® (the Virtual International Authority File), a worldwide project designed by a few national libraries and run by OCLC, which combines multiple name authority files with the goal “to lower the cost and increase the utility of library authority files by matching and linking widely-used authority files and making that information available on the Web.” It “clusters together the various forms of names for an entity” and has become “a major source for authority control” and is becoming the collective reference source at the international level’ (Bianchini, Bargioni, and di San Girolamo 2021, 2).

³¹⁶ ‘We’ve enhanced WorldShare Record Manager with the ability to look up WorldCat Entities within existing cataloging workflows and then add linked data identifiers to records. This functionality will also be added to other OCLC services in the future. WorldCat Entities URIs are now included in MARC records exported using OCLC cataloging tools. Individual exports of MARC records with linked data identifiers through WorldShare® Record Manager, Connexion®, the WorldCat Metadata EBPI, and Z39.50 protocol provide record-by-record access to linked data identifiers. And bulk output is available in WorldShare® Collection Manager.’ <https://www.oclc.org/go/en/publications/linked-data-the-future-of-library-cataloging.html> (Accessed 9 August 2024).

³¹⁷ ‘Application of LOD in library contexts is an active, current area of research. The application of LOD features to library collections and resources both increases the visibility of these resources on the web and provides end users with enhanced representations of primary sources, search results, and analytic information for research, especially within digital library special collections’ (Tian, Cole, and Yu 2021, 134).

³¹⁸ ‘The key point to enable these features is automated metadata reconciliation that maps bibliographic metadata from text strings to global Uniform Resource Identifiers (URIs). Successful reconciliation of name entities with VIAF authority records can enhance the user experience of digital library collections by accessing new and analytic information such as name variations for an author, titles associated with the author, and name forms in different languages’ (Tian, Cole, and Yu 2021, 133-134).

³¹⁹ ‘It is important to understand how VIAF views the clustering process. VIAF first matches each record in each source to all the records in all the other sources. Once those source-record-to-source-record links have been made, VIAF divides the records into clusters based on the links’ (Online Computer Library Center 2019, 5).

³²⁰ ‘Mixed names (mixed homonyms):

Tian, Timothy W. Cole and Karen Yu performed a comparative study of matches between VIAF and Library of Congress person Name Authority Files. The results indicated partial success and noted that the test failures were as a result of variable, inconsistent and incomplete MARC data. One sample returned only a 14% success rate.³²¹ OCLC, in collaboration with the British Library, is exploring using Wikidata³²² as an alternative

Mixed names often happen when titles that should be separately associated with different entities get all linked to one of these entities. This leads to merging unrelated entities, causing problems for everyone. If the record is known or suspected to be undifferentiated (mixed names), this needs to be indicated.

- Missing titles, dates. VIAF needs enough information to differentiate the entities you are describing. Personal names with no titles or dates associated with them (either from bibliographic records or from within the authority records) are difficult to match correctly; however, if the name appears unambiguous, matching is possible.
- Unusually encoded information. If an institution follows some local convention that is not typical (for example, how it differentiates between cities of the same name), please let OCLC know your conventions so it can try to utilize the information with profit.
- Differentiation using language-specific information. In personal names, this typically is in a MARC 21 \$c subfield as qualifier in the cataloging language (e.g., “Dramatist” in English vs. “Dramaturge” in French). For most matching, such information will be ignored because so much variation is seen.
- Duplicates. Duplicate records within a file cause problems because, in general, VIAF tries not to include more than one record from each source in a cluster. Duplicates often result in clusters being split. VIAF periodically reports records that look like duplicates. This can be useful, both in helping an institution to eliminate duplicates in its file and, at least as important, to identify problems in VIAF pulling together records that it should not, e.g., by ignoring some carefully coded information that an institution supplies but that OCLC is not aware of’ (Online Computer Library Center 2019, 4).

³²¹ ‘The match rate for the HAB sample (14.98 percent) is noticeably lower than others. One possible reason that may contribute to this low match rate is how many of the name entities in the HAB collection are formatted. They are formatted as name acronyms (which tend not to return matches) instead of full names. Another reason for match failure likely is that the lack of birth and/or death dates in many of the name entities returns too many results. As mentioned, match counts greater than 1 are not considered a successful match. Reason for the lack of dates in HAB name strings is unclear, but could be due in part to differences in metadata formatting and cataloging practices in Germany versus the US’ (Tian, Cole, and Yu 2021, 137). Numerous other instances of data cleaning issues are explored throughout the paper.

³²² ‘Wikidata stores stable and common information about entities, i.e., items and properties, and interlinks between different Wikimedia projects, in a form compliant with the RDF model (see <https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer>). Additionally, Wikidata uses triples and enriches them with qualifiers and references. Qualifiers allow adding specifications about the validity of a statement (start/end date, precision, obsolescence, series ordinal, etc.); references are fundamental to justify the data, i.e., to document the authority data creator's reason for choosing the name or form of name on which a controlled access point is based’ (Bianchini, Bargioni, and di San Girolamo 2021, 4).

model³²³ to the VIAF concept.³²⁴ The Wikidata route offers scope for including the contributions of non-academic actors because it is permissive in terms of access, and Wikidata has its own comparable name authority control standards.³²⁵ The work of OCLC and its international VIAF URI name authority aggregation, and the ongoing initiative to explore the permissive structures of Wikidata, suggests that EBP is a possible future pathway to invite family historians and genealogists into the community of archival contributors. This holds out the possibility of the enrichment of metadata on person name access points in archival records by the academic embracing of genealogical methodologies and the involvement of the general public.

³²³ 'The first attempt of cooperation between VIAF and Wikidata goes back to 2012, when Maximilian Klein and Alex Kyrios, Wikipedians in Residence at OCLC and the British Library, respectively, developed a project to integrate authority data from the VIAF with English Wikipedia biographical articles. The project successfully added authority data to hundreds of thousands of articles on the English Wikipedia, but above all showed that linking of data represents an opportunity for libraries to present their traditionally siloed data, such as catalogue and authority records, in more openly accessible web platforms. At the time, Wikidata was taking its first steps, but later authority data were successfully transferred from English Wikipedia to Wikidata' (Bianchini, Bargioni, and di San Girolamo 2021, 5).

³²⁴ 'VIAF contains personal name clusters, corporate name clusters, geographic name clusters, and work clusters, whereas Wikidata allows items to describe any kind of entity relevant in the universe of discourse of the users' data and irrespective of their bibliographic nature. Even if all kinds of VIAF clusters are relevant for bibliographic control, this study is limited to the analysis of personal name clusters in VIAF and of items having "instance of: human" (P31:QS) in Wikidata, because they are largely the most represented in VIAF and they can be directly compared ... The theoretical approach differs too, both as to the form of the names and as to identification function. In VIAF, preferred and variant forms of names for persons are based on national cataloguing codes. Because national codes are different, VIAF is needed and works as a neutral hub of all the national preferred forms. Cataloguing rules can assure uniformity and univocity to the forms of the names of the entities within a national catalogue but are quite complicated to be understood and used by users. In Ranganathan's words "the cataloguing conventions are on the surface quite contrary to what Mr. Everybody is familiar with." In contrast, preferred forms in Wikidata are based on the international principles of the convenience of the user and common usage. A clear example is the use of the direct form of name (Jane Doe) instead of the inverted form of name (Doe, Jane)' (Bianchini, Bargioni, and di San Girolamo 2021, 8 & 15).

³²⁵ '[E]ven if VIAF collects a huge amount of authoritative data and creates clusters of IDs, VIAF users can not always safely and continuously rely on them. Data flows just in one direction (from national libraries to VIAF), VIAF deletes and rebuilds clusters without giving priority to the stability of one cluster over another, and, after April 2020, VIAF no longer makes available to users a record of its changes. On the contrary, Wikidata data is always under strict control of any user, as its structure is designed to trace any minimum change to its data. Every single addition or deletion is documented, not just to easily recover eventual vandalism, but also to support any decision with clear evidence. Any stakeholder can exactly know if, how, when, and why data changed, in any moment' (Bianchini, Bargioni, and di San Girolamo 2021, 17).

4.4.2 International Standard Archival Authority Record for Corporate Bodies, Persons and Families, Second Edition 2004 ISAAR(CPF)

The metadata standard for archival authority records is ISAAR(CPF).³²⁶ The standard provides discipline in the catalogue metadata record making of archival records.³²⁷ Data is collected in archival metadata comprising archival record descriptions and their contexts. The standard enables precise identification of the originators of records and facilitates data interchangeability with other archives. This has the object of improving researcher access to the meaning and significance of archival records.³²⁸

GLAMS metadata frequently includes designated items called 'access points'. These are the points at which a researcher may choose to access the metadata record. An important access point is the name of the creator of the item (either a person or an organisation).³²⁹ Access point entries frequently follow either a national or archival authority naming rule, style or convention. These rules, styles or conventions are local in nature and, in order to facilitate interoperability standards, are carefully designed both to accommodate the need

³²⁶ 'Agents are entities that perform activities in the world. In the course of performing the activities the agents may generate or use record resources. The kinds of agents presented in RiC-CM include the entities represented in ISAAR(CPF): corporate bodies, persons, and families' (International Council on Archives 2023, 27).

³²⁷ Occasionally and confusingly, in archives the term 'record' is used to describe the item collected itself and also the data created by the archivist to describe the item.

³²⁸ 'The primary purpose, therefore, of this standard is to provide general rules for the standardization of archival descriptions of records creators and the context of records creation, thus enabling:

- access to archives and records based on the provision of descriptions of the context of records creation that are linked to descriptions of the often diverse and physically dispersed records themselves.
- understanding by users of the context underlying the creation and use of archives and records so that they can better interpret their meaning and significance.
- precise identification of records creators incorporating descriptions of relationships between different entities, especially documentation of administrative change within corporate bodies or personal change of circumstances in individuals and families; and the exchange of these descriptions between institutions, systems and/or networks' (International Council on Archives 2004, 1.10).

³²⁹ 'Archival authority records are similar to library authority records in as much as both forms of authority record need to support the creation of standardized access points in descriptions. The name of the creator of the unit of description is one of the most important of such access points. Access points may rely on the use of qualifiers that are deemed essential to clarify the identity of the entity thus named, so that accurate distinctions may be made between different entities that have the same or very similar names' (International Council on Archives 2004, 1.8).

for local conventions and at the same time to enable linked Open Data technologies to find and inter-relate them, across several languages and cultures.³³⁰ Standards in authority records also aim to discipline the recording variability of person names. Archival authority records are today often digitised and their structure is tightly controlled. Archival authority records can also be applied diffusely across the archive if many archival records are attributed to the same person.

While there are no international standards in the area of archival authority controls, there are often local standards (Chave and Sibille-de Grimoüard 2015). ISAAR(CPF) recognises the difficulty (or even impossibility) of creating international standards in archival authorities, and instead makes a general plea for their disciplinisation.³³¹ Interoperability is also a detailed and complex task, often too demanding for two archives to attempt alone. This is one of the reasons for the establishment of large multi-party cooperative efforts in the adoption of authority standards, which recognise both the considerable efforts that take place locally to ensure standard compatibility across the humanities and the need for interoperability and harmonisation.

A close examination of the ISAAR(CPF) standard reveals the access points relevant to EBP and provides an opportunity to assess the extent to which the NAI-UID system could be

³³⁰ 'Rules and conventions for standardizing access points may be developed nationally or separately for each language. Vocabularies and conventions to be used in creating or selecting the data content for these elements may also be developed nationally, or separately for each language' (International Council on Archives 2004, 4.9).

³³¹ 'Archival authority records are created primarily to document the context of records creation. To make this documentation useful it is necessary to link the authority records to descriptions of records. Archival authority records can also be linked to other relevant information resources. When such linkages are made it is important to describe the nature of the relationship, where known, between the corporate body, person or family and the linked resource' (International Council on Archives 2004, 6.0).

taken up using the extensibility feature of the standard. In Table 4.6, the text in the boxes indicated as **Comment** is mine.

Authority record
<p>The authorized form of name combined with other information elements that identify and describe the named entity and may also point to other related authority records.</p> <p>(International Council on Archives 2004, 10.0)</p>
<p>Comment: The standard permits that the form of the name should be that recorded in the archival authority record and the National Authority Record.</p>
Record
<p>Information in any fond or medium, created or received and maintained by an organization or person in the transaction of business or the conduct of affairs.</p> <p>(International Council on Archives 2004, 10.0)</p>
<p>Comment: Information may include numerous instances of EBPI (person names). The standard does not include these in its scope.</p>
Authorized form(s) of name
<p>Purpose: To create an authorized access point that uniquely identifies a corporate body, person or family.</p>

Rule: Record the standardized form of name for the entity being described in accordance with any relevant national or international conventions or rules applied by the agency that created the authority record. Use dates, place, jurisdiction, occupation, epithet and other qualifiers as appropriate to distinguish the authorized form of name from those of other entities with similar names. Specify separately in the Rules and/or conventions element which set of rules has been applied for this element. (International Council on Archives 2004, 5.1.2)

Comment: The application of the NAI-UID system here would simplify, regularise and make universally interoperable all person name access points. It would remove the need for prosopographical attributes to be used to clarify identification.

Authority record identifier

Purpose: To identify the authority record uniquely within the context in which it will be used.

Rule: Record a unique authority record identifier in accordance with local and/or national conventions. (International Council on Archives 2004, 5.4.1)

Comment: The standard facilitates the NAI-UID system.

Sources³³²

³³² 'Where a number of repositories hold records from a given source they can more easily share or link contextual information about this source if it has been maintained in a standardized manner. Such standardization is of particular international benefit when the sharing or linking of contextual information is likely to cross national boundaries. The multinational character of past and present record keeping creates the incentive for international standardization which will support the exchange of contextual information. For

Purpose: To identify the Sources consulted in creating the authority record.
Rule: Record the Sources consulted in establishing the authority record.
Examples:
HMC, Principal Family and Estate Collections: Family Names L-W, 1999
Complete Peerage, 1936
Burkes Peerage, 1970
Complete Baronetage, vol 5, 1906
United Kingdom, The National Archives: Historical Manuscripts Commission (International Council on Archives 2004, 5.4.8)
Comment: The NAI-UID system affords the best primary source. Other sources should only be used when the NAI-UID system does not produce a satisfactory entry.
Internal structures/genealogy
Purpose: To describe and/or represent the internal administrative structure(s) of a corporate body or the genealogy of a family.
Rules: Describe the internal structure of a corporate body and the dates of any changes to that structure that are significant to the understanding of the way that corporate body conducted its affairs (e.g. by means of dated organization charts). Describe the genealogy

example, processes such as colonialization, immigration and trade have contributed to the multinational character of recordkeeping' (International Council on Archives 2004, 1.7).

of a family (e.g. by means of a family tree) in a way that demonstrates the inter-relationships of its members with covering dates. (International Council on Archives 2004, 5.2.7)

Comment: With respect to persons, the NAI-UID system and conventional genealogical databases should be used to provide unambiguous primary data.

Category of relationship

Purpose: To identify the general category of relationship between the entity being described and another corporate body, person or family.

Rule: Record a general category into which the relationship being described falls. Use general categories prescribed by national rules and/or conventions or one of the following four categories. Record in the Rules and/or conventions element (5.4.3) any classification scheme used as a source of controlled vocabulary terms to describe the relationship.

Family: In a family a person may have a wide circle of relationships with other members of the family and with the family as an entity. Where the genealogical structure of the family is complex it may be appropriate to create separate authority records for each member and link them to parent(s), spouse(s) and child(ren). Alternatively, this information may be recorded in the Internal structures/Genealogy element. (International Council on Archives 2004, 5.3.2)

Comment: The use of the NAI-UID system removes the need for an archive to create person records if those persons are not present in the archive records.
Names/identifiers of related corporate bodies persons or families
Purpose: To indicate the names and any unique identifiers of related entities and to support linkages to the authority records for related corporate bodies
Rule: Record the authorized form of name and any relevant unique identifiers including the authority record identifier for the related entity. (International Council on Archives 2004, 5.3.1)
Comment: The NAI-UID system provides a single national system replacing many separate systems created by each archive.

Table 4.6 Extracts in tabular form (International Council on Archives 2004, 1.7)

The ISAAR(CPF) standard is a core component of many international archival authority control systems, and the detailed examination here shows that the EBP and NAI-UID system is capable of inclusion within the standard through its extensibility feature with minimum impact on other parts of the standard. The NAI-UID system simplifies archival record making, lessens the need for using attributes in the standard to reinforce identification, lessens the burden of interpretation on the archivist, and makes the standard interoperable at the level of person related access points. Furthermore, it affords researchers improved access to materials and their relationships.

4.4.3 Describing Archives: A Content Standard (DACS)

The Society of American Archivists made the first release of DACS in 2005,³³³ and the standard underwent a major revision in 2013. From its beginnings DACS was designed to have multiple uses, and the 2013 revision recognised the growing interest in the convergence of GLAMS descriptive standards, especially through the widespread adoption of RDA.³³⁴ Nevertheless, DACS is not a stand-alone solution to archival standards, because it relies heavily on several related ‘Companion Standards’.³³⁵ Beginning in 2016 the DACS drafting team have been working with the RiC team to bring the two standards into close alignment.³³⁶

Of particular importance to the work of this thesis is how the DACS standard directs metadata entries for person names. Names can be recorded at nominal access points in several places in DACS, and freedom is granted locally to determine how and where person names are entered.³³⁷ A selection of suitable NAR DACS standard access point definitions

³³³ ‘The Society of American Archivists adopted Describing Archives: A Content Standard (DACS) as the official content standard of the U.S. archival community in 2005. DACS was designed to be used to create a variety of archival descriptions, including finding aids and catalog records’ (Society of American Archivists 2020, vi).

³³⁴ ‘In 2013, following a call from the Council of the Society of American Archivists and after soliciting feedback from the community, DACS underwent a major revision. The revisions addressed the growing convergence between archival, museum, and library descriptive standards—particularly the promulgation and adoption of RDA’ (Society of American Archivists 2020, vi).

³³⁵ ‘As a content standard, DACS is part of an ecosystem of interrelated and, in some cases, interdependent standards which support the process of archival description. Sometimes referred to as “companion standards,” these include structure standards, other content standards, and communication standards. DACS relies on two international content standards for archival description: International Standard Archival Description-General (ISAD(G) and the International Standard Archival Authority Record for Corporate Bodies, Persons, and Families (ISAAR(CPF)). All of the data elements of ISAD(G) and ISAAR(CPF) are incorporated into DACS’ (Society of American Archivists 2020, VI & VII).

³³⁶ ‘Following the draft release of its conceptual model in 2016, DACS and Records in Contexts (RiC) are now entering a period of coevolution’ (Society of American Archivists 2020, vii).

³³⁷ ‘It is a local decision as to which names, terms, and concepts found in a description will be included as formal access points, but repositories should provide them in all types of descriptions. Such indexing becomes

appears at Appendix 5. The DACS schema allows person names to be allocated a URI from an Archival Authority Index (AAI-UID). For the purposes of EBP, an AAI-UID can be easily referenced to an NAI-UID. This would allow a researcher to validate person names through the archive and its sources to the NAI-UID record based on BMD records, thus facilitating relationship searches at research project, archive sources and national authority levels. The archivist is, by NACO convention, free to assemble person name data from any ‘reliable Sources’,³³⁸ as long as a consistent policy for person names is applied across the archive. From a researcher perspective it would be preferable that incidences of person names in sources should be referenced to a national authority system to maximise relationship searches and facilitate research collaboration and data interoperability.

4.4.4 Records in Contexts Conceptual Model (RiC-CM)

RiC-CM³³⁹ was adopted by the ICA in 2016, when it changed its approach to archival description because its traditional fonds-based structure (a rigid hierarchical structure) was found to be inadequate, since it ‘often does not reflect the social and material complexity of

increasingly important as archivists make encoded finding aids and digital content available to end users through a variety of repository-based and consortial online resource discovery tools’ (Society of American Archivists 2020, xix).

³³⁸ ‘Assemble the information from reliable sources, such as the materials themselves and reference works. Establish a consistent policy regarding the content, form, and placement of citation of sources and quotations’ (Society of American Archivists 2020, 30).

³³⁹ ‘RiC-CM is a high-level conceptual model that focuses on intellectually identifying and describing records, the people that created and use(d) them, and the activities pursued by the people that the records both facilitate and document. RiC-CM covers all of the essential content of the four existing International Council on Archives (ICA) description standards: General International Standard Archival Description (ISAD(G))2; International Standard Archival Authority Records for Corporate Bodies, Persons, and Families (ISAAR(CPF)); International Standard for Describing Functions (ISDF); and International Standard for Describing Institutions with Archival Holdings (ISDIAH). RiC-CM replaces these four standards in one overarching standard. It incorporates from them the core descriptive entities, the properties or attributes of these entities, and the essential relations between the entities’ (International Council on Archives 2023, 1).

the origins of records'.³⁴⁰ The new approach taken by the ICA with RiC-CM was to replace the fonds hierarchical structure with a graph based relational structure which allows each entity to have multiple and complex relationships.³⁴¹ RiC-CM provides a foundation for producing high-quality knowledge graphs describing records and their contexts. RiC-O is a formal implementation of RiC-CM that defines the vocabulary and rules for representing archival descriptions as RDF graphs (see Figure 4.7).

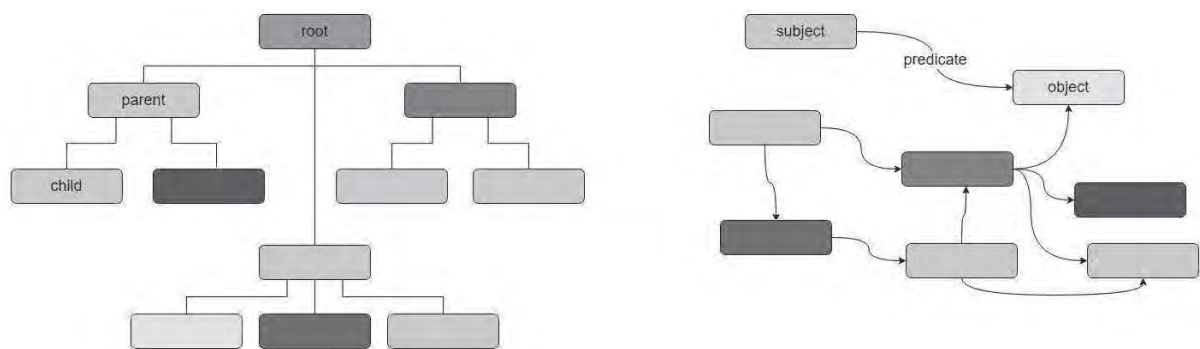


Figure 4.7 Left: Representation of data in a hierarchical structure like XML or other markup-language. Right: Representation of data in triples that results in a graph structure (International Council on Archives 2023, 6)

³⁴⁰ 'RiC-CM recognizes that provenance is much more complex, that records originate and continue to exist within a complex network of dynamic relations with other records, activities, persons, and groups' (International Council on Archives 2023, 7). 'How records are represented by archival description has also been a contentious topic in archival science. The traditional approach has been to the fonds, according to the principles of provenance and original order. The several international standards for archival description published by the International Council on Archives (ICA) throughout the 1990s have prioritized collection-level descriptions of records, according to a hierarchical structure from the general to the particular, from the fonds to the item. The ICA has since then changed its shift and, in its latest standard Records in Contexts, it acknowledges that this focus on the person or group that has accumulated a body of records "often does not reflect the social and material complexity of the origins of records"' (International Council on Archives 2016, 5). Quoted in Pacheco, Da Silva and De Freitas (2023, 636). See also International Council on Archives (2023, 6).

³⁴¹ 'RiC-CM models what may be described as "multidimensional description." Rather than a hierarchy, the description may take the form of a graph or network. Modelling description as a graph accommodates the single, fonds-based, multilevel description modelled in ISAD(G), but also enables a more open description of the often-complex and mixed provenance of records found in a fonds. The model makes it possible, using various relations between record resources and agents or activities, to describe sets of records with complex origination, for example, a record series that documents one activity that is performed serially by a succession of different groups, and at the same time, situate the series within the fonds of the different groups that serially had responsibility for the activity' (International Council on Archives 2023, 11).

In the RiC-CM, persons are classified as Agents (RiC-E07), and within the class of Agent each person name is recorded (RiC-E09) and can then be related to family members (RiC-E10) (Figure 4.8).

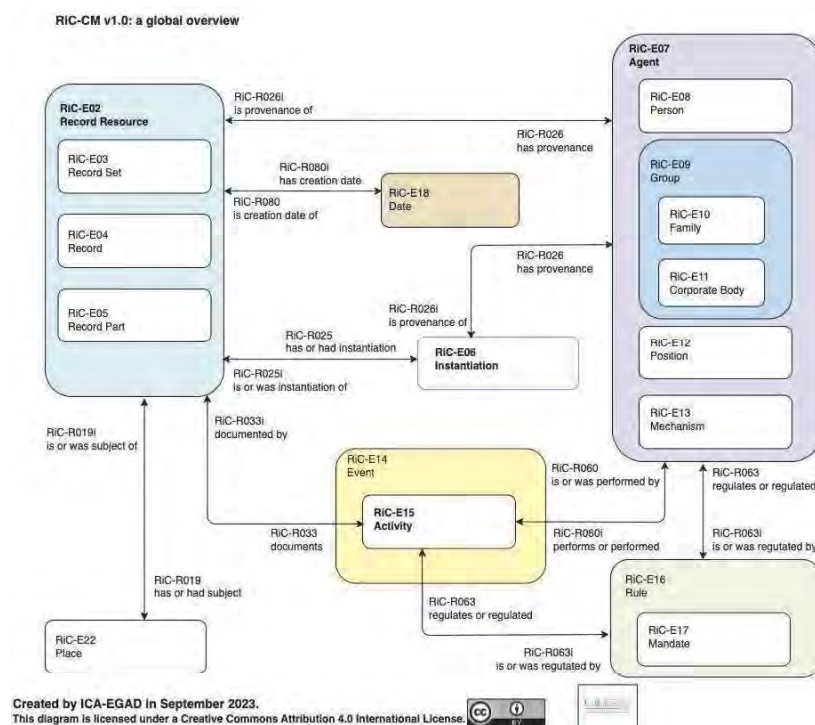


Figure 4.8 Overview of the RiC Conceptual Model – note RiC E07 Agent class (top right) (International Council on Archives 2023, 18)

4.4.5 Comité International pour la Documentation of the International Council of Museums (CIDOC CRM)

CIDOC CRM is now an international standard reference model (ISO 21127:2006). It was first published in 1998,³⁴² replacing an older entity relationship model that had become cumbersome and no longer fit for purpose.³⁴³ In developing the new standard the Special Interest Group (SIG) tasked with its development recognised that it should have a unifying capability across all of GLAMS.³⁴⁴ This resulted in the SIG adopting a broad and ambitious object of interest, ‘the empirically investigable human past but several with regards to its approaches’ (Bruseker, Carboni, and Guillem 2017, 95). This wide scope is one of several defining features of the CRM (see Figure 4.9).³⁴⁵

³⁴² ‘The museum community published the first version of its conceptual model at the same time as the library world: in 1998, when FRBR appeared, the Comité International pour la Documentation (CIDOC) of the International Council of Museums, published the CIDOC CRM – Conceptual Reference Model’ (Bruseker, Carboni, and Guillem 2017, 251).

³⁴³ ‘Until 1998 the CIDOC organisation (the documentation wing of the International Council of Museums, ICOM) had maintained a traditional Entity Relationship model (E-R model) – a modelling system used in the design of relational database systems) of the cultural heritage domain largely derived from work by the Smithsonian Institute. However, the E-R model exposed some major flaws. Its lack of flexibility and semantic capability meant that the model continually expanded to reflect new information requirements and variations, but consequently became too complex; as a result additional areas of practice were increasingly difficult to represent properly and the model became unmaintainable’ (Oldman 2014, 2).

³⁴⁴ ‘The disciplines of archaeology, conservation, museology, library studies, archives, and so on, should not operate in a vacuum from each other’s research results. The outcomes of the one, assuming they all refer to the same objective domain of discourse, have implications on the other which require assimilation and integration into the overall view of affairs, potentially initiating knowledge revisions or new conclusions based on new information revealed by techniques, or methods of study not available in one’s own home disciplines’ (Bruseker, Carboni, and Guillem 2017, 95).

³⁴⁵ ‘Data coming from the cultural heritage community comes in many shapes and sizes. Born from different disciplines, techniques, traditions, positions, and technologies, the data generated by the many different specialisations that fall under this rubric come in an impressive array of forms. Considered together the collective output of this community forms a latent pool of information with the capacity, when integrated, to support potential knowledge generation relative to any period, geographic location, or aspect of human activity in the past even when, characteristically, based on sparse datasets’ (Bruseker, Carboni, and Guillem 2017, 93-4).

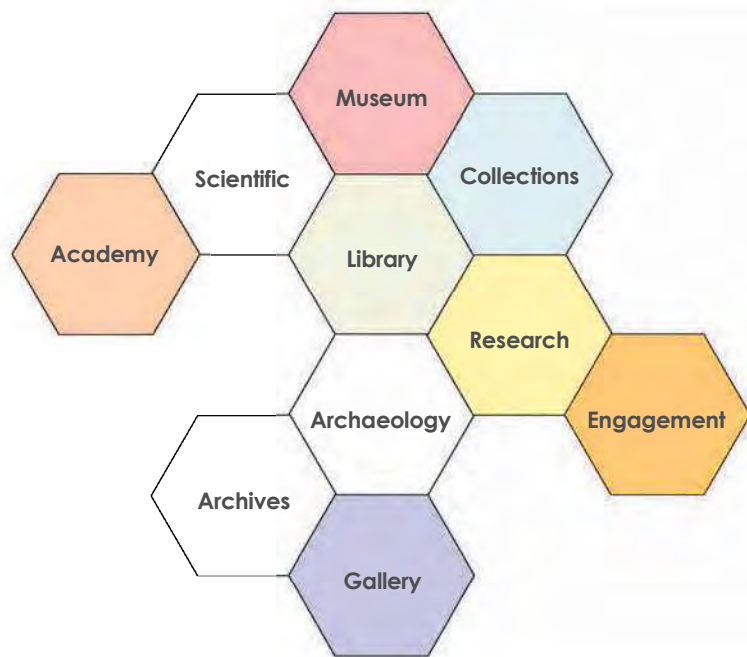


Figure 4.9 'The primary role of the CIDOC CRM is to serve as a basis for mediation of cultural heritage information and thereby provide the semantic "glue" needed to transform today's disparate, localised information sources into a coherent and valuable global resource' (Oldman 2014)

The CRM is RDF based and a typical element within it might look as follows:

Subject: [HTTP://www.digbib.org/Franz_Kafka_1883/Das_Schloss](http://www.digbib.org/Franz_Kafka_1883/Das_Schloss)

Predicate: [HTTP://www.cidoc-crm.org/rdfs/cidoc-crm#P14_carried_out_by](http://www.cidoc-crm.org/rdfs/cidoc-crm#P14_carried_out_by)

Object: [HTTP://www.viaf.org/viaf/56611857](http://www.viaf.org/viaf/56611857)

In this example the subject is an entry for the publication *Das Schloss* in the page for Franz Kafka in the German Free Bibliography. The predicate is the link to the CIDOC CRM RDFS property declarations P14 – carried out by (performed). The object is the link to the VIAF entry for Franz Kafka (a URI), (Oldman, Doerr, and Gradmann 2015, 253).

The CRM also captures time and space coordinates relative to persons, objects and events. This approach satisfies both the need for museums to be able to place objects in time and

space, and the need to be able to capture changes as object attributes change over their lifetimes.³⁴⁶ CIDOC CRM in this regard is unique, and as well as providing an important CRM for museums, it offers a complementary standard to GLAMS in general, especially for archival collections that in their nature persist over time.

Because CIDOC CRM functions over two dimensions, space and time, an example taken from the CIDOC manual is the easiest way to show the new affordance that the CRM offers. The example ‘Winkelmann seeing Laocoön’ is taken from the CIDOC CRM SIGs Definition of the CIDOC CRM Conceptual Reference Model 2021 (Bekiari 2021). The event is represented as an Entity Relationship Diagram (ERD) as in Figure 4.10.

³⁴⁶ ‘The CRM does not stop with expressing entities and relationships between static things characteristic of the bibliographic universe, but pays attention to processes which happen in time and, of course, space, and therefore enables description of complex dynamic objects, and the lifecycle of objects. That is why the CRM is considered to be event driven; it defines entities or classes such as Temporal Entity, Persistent Item or Person, but also Event, Activity, Beginning of Existence, Birth, Dissolution, and Death. In fact, the latter classes are considered to be the basic ones for expressing explicitly dynamic views of things in the former classes. In this respect, and because of its comprehensive scope for describing the complex “population” of what is basically the museum universe, it attracted information system designers to use it as a reference model for services that have as their aim the integration of heterogeneous information from across the range of heritage and memory institutions’ (Bruseker, Carboni, and Guillem 2017, 252-253).

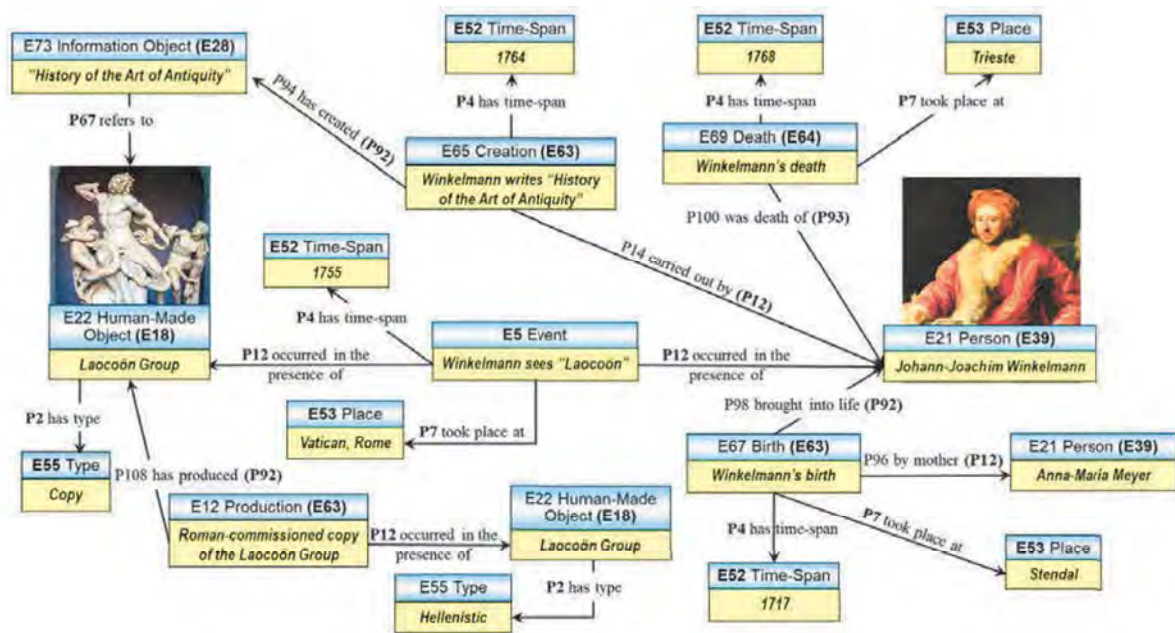


Figure 4.10 CIDOC CRM encoding example (Winkelmann seeing Laocoön) (Bekiari 2021, 36)

The ERD displays the relationships embedded in this text:

Johann-Joachim Winkelmann (a German Scholar) has seen the so-called Laocoön Group in 1755 in the Vatican in Rome (at display in the Cortile del Belvedere). He described his impressions in 1764 in his 'History of the Art of Antiquity', (being the first to articulate the difference between Greek, Greco-Roman and Roman art, characterizing Greek art with the famous words '... noble simplicity, silent grandeur'). The sculpture, in Hellenistic 'Pergamene baroque' style (Bieber 1961, Brilliant 2000) is widely assumed to be a copy, made between 27 BC and 68 AD (following a Roman commission) from a Greek (no more extant) original. Johann Joachim Winkelmann was born 1717 as child of Martin Winkelmann and Anna-Maria Meyer and died in 1768 in Trieste.³⁴⁷ (Bekiari 2021, 36)

³⁴⁷ Figure 4.10 'presents a semantic graph of this event, as described above, using CIDOC CRM concepts. The facts in parentheses above are omitted for better clarity. Instances of classes are represented by informative labels instead of identifiers, in boxes showing the class label above the instance label. Properties are represented as arrows with the property label attached. After class labels and property labels, we show in parenthesis the identifiers of the respective superclasses and superproperties from [Figure 4.10], in order to demonstrate that the story can be represented and queried with these concepts only. It also shows how concept specialization increases expressiveness without losing genericity. It is noteworthy that the transfer of information from the Greek original, to the copy, to the mind of Winkelmann and into his writings can be solely understood by this chain of things being present in different meetings. Note also that the degree to which a fact is believed to be real does not affect the choice of CIDOC CRM concepts for description of the fact, nor the reality concept underlying the Model' (Bekiari 2021, 36).

The CIDOC CRM can also display the same information in time and space by using the axis of a chart to represent the two dimensions (see Figure 4.11).

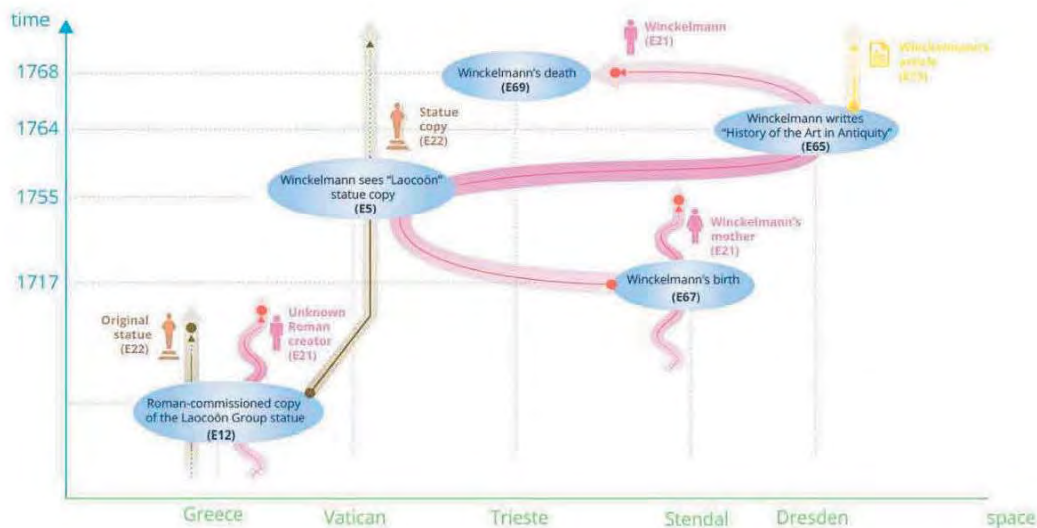


Figure 4.11 Symbolic representation of 'Winkelmann seeing Laocoon' as an evolution in space and time (Bekiari 2021, 37)

In the 'Winkelmann seeing Laocoön' example it can be assumed that a suitable AAI (or equivalent selected by the archivist) has been used to compile the underlying metadata in the archival 'core' and 'detail' levels. The CRM collects these under the entity classes 'Actors', 'Events' and 'Objects'. These acquire authority from information entered in 'Extracted knowledge data' and 'Background knowledge/Authorities' (RDF), and these are sourced in turn from 'Sources and metadata' (XML/RDF). (See Figure 4.12.)

The CIDOC CRM-Application Repository Indexing

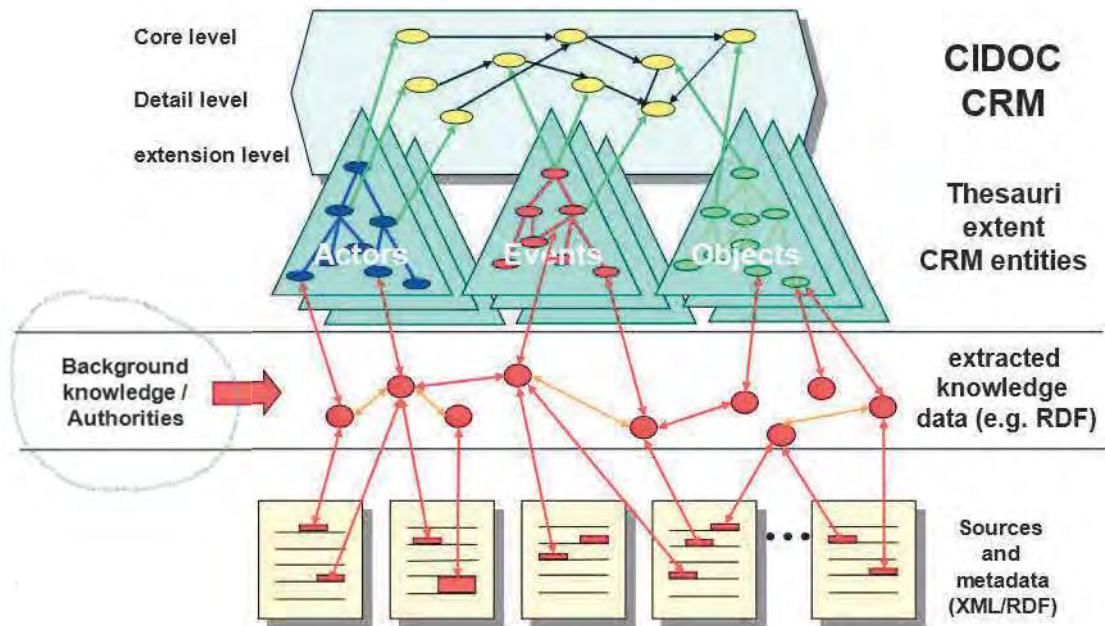


Figure 4.12 Authorities come from 'Background knowledge' in the CIDOC CRM (Stead 2008)

The detailed examination of the RiC-CM in Section 4.2.2 and a crosswalks study of RiC and CIDOC CRM³⁴⁸ suggests that, while CIDOC CRM also permits interpretation and determination at the archivist's discretion in the selection of appropriate sources, like in RiC-CM multiple entries are permissible and URIs are acceptable (see the highlighted text in Table 4.7).

³⁴⁸ 'A CIDOC CRM-compatible knowledge base (KB) (Meghini & Doerr 2018) is an instance of E73 Information Object in the CIDOC CRM. It contains (data structures that encode) formal statements representing propositions believed to be true in a reality by an observer. These statements use appellations (e.g., <http://id.loc.gov/authorities/names/n790660056>) of ontological particulars and of CRM concepts (e.g., P 1 00i died in). Thereby users, in their capacity of having real-world knowledge and cognition, may be able to relate these statements to the propositions they are meant to characterize, and be able to reason and research about their validity' (Bekiari 2021, 26).

E21 Person
Subclass of:
E20 Biological Object
E39 Actor
<p>Scope Note: This class comprises real persons who live or are assumed to have lived. Legendary figures that may have existed, such as Ulysses and King Arthur, fall into this class if the documentation refers to them as historical figures. In cases where doubt exists as to whether several persons are in fact identical, multiple instances can be created and linked to indicate their relationship. The CIDOC CRM does not propose a specific form to support reasoning about possible identity. In a bibliographic context, a name presented following the conventions usually employed for personal names will be assumed to correspond to an actual real person (an instance of E21 Person), unless evidence is available to indicate that this is not the case. The fact that a persona may erroneously be classified as an instance of E21 Person does not imply that the concept comprises personae.</p>
Examples
<ul style="list-style-type: none"> • Tut-Ankh-Amun (Edwards and Boltin, 1979)
<ul style="list-style-type: none"> • Nelson Mandela (Brown and Hort, 2006)

Table 4.7 Entity 21 Person in the CIDOC CRM (Bekiari 2021, 73-74)

CIDOC CRM requires multiple person names to be associated with a given object over time. The CRM also has the desire to achieve seamless interoperability with RiC. Because CIDOC CRM is a recent affordance in the humanities, it agrees with the ISAAR(CPF) standard recommendation to use URIs wherever possible in person name access points. For these reasons, the CRM is well able to adopt the EBP NAI-UID system proposed by this thesis. As was found with ISAAR(CPF) and RiC, the CIDOC CRM would benefit by adopting the NAI-UID system because it could then be applied universally across all GLAMS disciplines, improving interoperability.

4.5 Chapter summary

This chapter has taken a detailed view of GLAMS and how and where EBPI fits into GLAMS standards and ontologies. To clarify the object of EBP with regard to GLAMS, one section described EBPI as it is found in GLAMS, and another how EBPI in GLAMS can become more accessible and useful in research, especially when it is enhanced through LOD affordances.

The leading institutions in GLAMS were discussed together with their roles in standards building. Important new developments in metadata in the areas of UBC and AAC in relation to NAR were discussed to show the direction of travel for standards in GLAMS. The role of NACO (which is a standard in the US only but is used as a model internationally) in NAR in developing guidance on the establishment of person and family name indexing and authority structures across GLAMS was set out.

Finally, a detailed analysis was performed on leading affordances in GLAMS standards and name authority affordances in libraries, archives and museums. The chapter has shown that digitised finding aids are a good bridge to the Records and the informations contained in them that researchers are interested in, and that they could be improved by adopting the EBP system, which would extend the findability of EBP in Records and systematise the linking of EBP to the NAI system, thus improving provenance tracking and data control (fixity and affixedness). It was noted that Special Collections are probably not large or rich enough to develop specialist affordances at the GLAMS level, but instead would benefit from the general affordances improvements. However, Special Collections figure more prominently in EBP in the area of research affordance, which is the subject of Chapter 5.

Chapter 5 EBP in recent research projects

5.1 Introduction

How successfully have recent large-scale research projects into Past Human Lives used Evidence Based Prosopography? EBP is at the centre of much new research focused on the interconnectedness and relationships inherent in the study of PHL, and this chapter considers a small selection of recent high profile EBP research projects, whose development was contemporaneous with this project's time-frame. They each have features that utilise EBP, and they illustrate recent attempts to build datasets and/or research aids that could be used to explore prosopographical data. It will be shown that each of them could have been improved by adopting EBP and the NAI-UID system.

They are all (albeit larger-scale and more sophisticated) examples of EBP research projects similar to this project's P7 Case Studies, which are discussed in the following chapter. The P7 Case Studies follow similar concepts and use similar technologies and datasets as the five projects considered here. They each use arrays of separate technologies which they combine to form integrated suites of research tools, and in their technical aspects they are a more expensive and complex equivalent of this project's Human Data Digital Toolkit (HDDT). In this sense this chapter provides a context for the P7 project discussed in Chapter 6 and Chapters 3 and 4 have provided the critical background to the analysis of the research projects considered in this chapter. The research projects examined are:

- Social Networks and Archival Contexts (SNAC)
- The Cambridge Group's I-CeM project

- ResearchSpace
- Traces through Time (TTT)
- The Golden Agents project

The projects considered here come from the EU, the US and the UK. They have all been chosen in part because they exhibit the influences of different national approaches to the provision of research project infrastructure support – ‘top down’ in the EU and ‘bottom up’ in the US (Kaltenbrunner 2017), with the UK’s as yet undeveloped future direction probably a mixture of both approaches. Chapter 4a took an in-depth look at metadata found in GLAMS records and questioned the suitability of GLAMS metadata itself as a surrogate for EBPI contained in sources. Each project here relies heavily on combining a variety of pre-existing and also differently structured datasets formed from a mixture of GLAMS metadata and primary sources. The projects therefore offer an opportunity to critique the use of GLAMS metadata for research into PHL in practice and compare it to the use of data that represents primary sources. This chapter will show how the issues and concerns identified in Chapters 3 and 4 arise in practice at the research level (even in very expensive high profile projects). The balance of reliance on Representative Data (primary) and metadata (secondary) sources severely conditions and restricts the relative success of each of the projects. Perhaps more significantly, this chapter calls into question the world of DH datasets, and how lack of clarity in understanding and managing the complexities of both Representative Data and metadata unavoidably gives rise to problems in dataset reliability and interoperability.

To make this clear, this chapter shows how affixedness and provenance are both essential and determining features of good research into PHL. Each project has grappled to various

extents with affixedness and provenance, especially when dealing with messy or incomplete data at source. It is shown here that the application of EBP principles and practices, and the adoption of the NAI-UID system, would increase the research value of the DH projects considered and, by extension, would improve the affixedness and provenance of all DH projects that focus on PHL, because these projects must all rely on prosopographical data found in sources.

Analysing these examples in the light of EBP shows that it is necessary to identify prosopographical information separately from where it is today embedded, mixed in with archivists' expert opinions in GLAMS data. Only EBP data can be used to build a national system of structured, uniform and consistent Linked Open Data (LOD) for PHL, importantly leaving non-evidence based and other quasi-prosopographical data to be accommodated solely at the local project level.

Indiscriminate mixing of primary and secondary sources in data science is to be avoided at all costs, if in the world of data science, data is to be relied on as a surrogate for the physical item it digitally represents. It is a problem similar to that of confusing primary and secondary sources in the traditional practice of historical research (which academia has universally enforced as a practice discipline). Indiscriminate mixing of primary and secondary sources in data science should be avoided at all costs, if in the world of data science data is to be relied on as a surrogate for the physical item it digitally represents.

This will over time become even more important if due to primary source losses digital representations *must* be relied on in the future. Future EBPD project ontologies must recognise and separate representative EBPD from related secondary sources (especially GLAMS catalogue metadata). Future research data integrity relies on clearly identifiable,

consistent, comparable and verifiable data, which in the field of PHL data is EBP. Carefully and clearly separating EBP data from other kinds of secondary research data is essential to the future integrity of DH research into PHL. Without the systematisation of affixedness and provenance in unambiguously identifiable EBP, the future interoperability of research datasets and thereby the systematic accumulation of knowledge of PHL is in doubt. The sure accumulation and linking of traversable and combinable data on PHL cannot be robustly achieved unless the focus of DH moves away from tools and technologies and towards a science based focus on the richness, complexity and messiness of data itself.

This chapter also recognises, but does not consider, the efforts made by others in DH and related disciplines to enable researchers to better explore the digital representations of historical documents (such as Corpus Linguistics and image analysis). The application of data science and the critical questioning of data and data philosophies, data systems and data applications are a common and shared enterprise across all of DH, but this thesis is focused on GLAMS and the research into Records held in GLAMS institutions.

Each of the five representative projects is an exemplar of affordances in the academic study of person-centred LOD, and each of them exploits EBPI extracted from sources taken from several different, but related, archives. The order in which they appear in this chapter simply enables the arguments made in this thesis to be logically considered.³⁴⁹

This chapter begins with a consideration of the Social Networks and Archival Context (SNAC) project (Section 5.1). The problems of relying on archival metadata of variable quality and the questionable use of the Friend of a Friend (FOAF) ontology to generate LOD, together

³⁴⁹ The projects considered are compared one to another merely to enable a discussion of the themes of this thesis. They are not ranked critically, each of them addresses the challenge of LOD in GLAMS in its own way and each has strengths and weaknesses.

with a lack of user intervention provisions, illustrate the issues that can arise in LOD affordances. The Cambridge Group for the History of Population and Social Structure (The Cambridge Group) and its I-CeM project are considered in Section 5.2. This project demonstrates the utility of EBP, especially BMD and census data. While recognising that this data source also has its challenges, the project proves that EBPD is a better alternative to archival metadata when used in LOD affordances. The Traces through Time project is considered in Section 5.3. Here EBP data has been chosen by the TTT project to link person name appearances across many datasets. ‘Fuzzy data’ techniques were deployed to help overcome the problems of name matching identified by The Cambridge Group. The problems of integrating EBP LOD taken from several digital catalogue and finding aids is made clear. ResearchSpace is considered in Section 5.4. It provides a detailed evaluation of an EBP LOD affordance through the related Linked Conservation Data project.

ResearchSpace provides a desktop environment on which data items (assets) can be arranged and interlinked through relationships. While ResearchSpace worked well as a core project tool, the Linked Conservation Data project which used ResearchSpace exposed concerns of a generic nature. These derived from the unavoidable complexity of digital research affordances of this kind, and the problems in coping with multiple data sources, each with its different ontologies and vocabularies. Finally, in Section 5.5 the Golden Agents project is considered because it appears to have successfully addressed all of the concerns raised from the analysis of the previous four projects in terms of its use of EBP. It has achieved its success largely because the project team adopted an information science approach from the outset, seeking only EBPD (not mixing EBP with metadata) and managing EBPD with great attention given to the key concerns of affixedness and provenance.

Nevertheless, even the Golden Agents project could still be improved, specifically through

the adoption in the Netherlands of a National Authority Index (NAI) system based on EBP to which the project could link and then bridge to other digital affordances in the Netherlands.

5.2 Social Networks and Archival Context (SNAC)

The SNAC project began in 2010 with funding from the US National Endowment for the Humanities, with the objective to find and match various biographical data elements embedded in archival metadata records at several institutions.³⁵⁰ The project made a research tool to house the data and it continues to expand by adding more data.³⁵¹ After an initial development phase, in 2015 the Andrew W. Mellon Foundation sponsored the second phase of the project with additional funding from the US Institute for Museum and Library Services. The development team then expanded to form a 'cooperative' of related institutions and include more datasets.³⁵²

³⁵⁰ 'SNAC (Social Networks and Archival Context) is a free, online resource that helps users discover biographical and historical information about persons, families, and organizations that created or are documented in historical resources (primary source documents) and their connections to one another. Users can locate archival collections and related resources held at cultural heritage institutions around the world. SNAC is an international cooperative including, but not limited to, archives, libraries, and museums, which is working to build a corpus of reliable descriptions of people, families, and organizations that link to and provide a contextual understanding of historical records.' <https://portal.snaccooperative.org/about> (Accessed 19 August 2024).

³⁵¹ 'In 2010, with funding from the U.S. National Endowment for the Humanities, SNAC began to explore the value of extracting the biographical and historical data about the individuals who created or are documented in archival records from online record descriptions. This data was then assembled into a collection of descriptions showing the individuals, families, and organizations and how they are interrelated with one another and with the archival resources that document their lives. SNAC next used the collection of descriptions to build a History Research Tool that 1) integrates and simplifies access to the dispersed resources and 2) provides unprecedented access to the biographical-historical contexts of the people documented in the resources, including the social-professional-intellectual networks within which they lived.' <https://portal.snaccooperative.org/node/356> (Accessed 12 October 2024).

³⁵² 'In 2015, the core team, which included the University of Virginia, Institute for Advanced Technology in the Humanities; the University of California, Berkeley School of Information; and the California Digital Library (part of the University of California), transformed this research into an international cooperative hosted by the U.S. National Archives and Records Administration (NARA). Thus the cooperative program was begun, focusing on the development of a governance infrastructure, technical infrastructure and sustainability, and the end-user experience.' <https://portal.snaccooperative.org/node/356> (Accessed 12 October 2024).

In the SNAC project, which links prosopographical-type data across several archives, the linking of metadata between contributing datasets was made using the links 'Creator Name', 'Related Name' and other 'Access Point' fields found in the various metadata catalogues and authority records of the contributing archives. As Chapter 4 has shown, such data has been compiled over many years by many archivists and to a variety of evolving standards, which were often at variance with each other, especially at the beginning of digitisation in GLAMS, but are now increasingly convergent. Convergence, as has been shown, comes at a cost: data mapping is a complex process requiring accommodations to be made at least for some items in virtually all large-scale data merging activities. There is considerable variability in the data quality of legacy data records. The project also relies on linking archival records (metadata) where those records are a mixture of prosopographical and other informations. These aspects of the SNAC project give rise to concern for researchers. These concerns are partly outweighed by the seeming utility of SNAC for archivists. From a general point of view, looking for prosopographical data is difficult because it can be hard to find, often taking years of researcher time, and in this regard digitisation is a significant and valuable even if imperfect aid to research.

The SNAC Cooperative research tool achieved its mature form in 2015.³⁵³ It offers a significant contribution to improve resource finding by digitally separating the descriptions of creators from the descriptions of the archival resources themselves, making digital

³⁵³ 'At the time of writing, SNAC is currently in its second phase (2012–2014), which is funded by the Andrew W. Mellon Foundation. In this phase, we are vastly expanding the quantity and diversity of the source data, thereby creating an even bigger corpus of EAC-CPF records and extending its research potential. We are also continuing to improve the prototype access interface, by conducting user testing and adjusting the design to better meet researcher need' (Pitti et al. 2015, 78).

management of data descriptions more economical, flexible and useful.³⁵⁴ An interface tool was built to join records at the access points. In SNAC it remains the case that, like the contributing datasets, SNAC points to the archival records underpinning the metadata.³⁵⁵

The Cooperative had a vision to ‘ultimately benefit the scholarly users of the primary resources, by offering unprecedented efficiencies and rendering explicit the social-document network that currently lies hidden’ and to ‘set the stage for a cooperative program for maintaining names of creators of archival materials, via the Encoded Archival Context—Corporate Bodies, Persons, and Families (EAC-CPF) standard’ (Pitti et al. 2015, 78 and 95).

The objectives of SNAC are to:

- extract the names, social-professional, and resource relations, and biographical-historical data from existing, dispersed descriptions of archival resources—primarily in the form of existing EAD finding aids, MARCXML records, and original archival authority records.
- format the data into standardized identity records, leveraging the EAC-CPF standard.
- match the resulting identity records against one another and against data in the Virtual International Authority File, combining records that identify the same entity, to produce a set of unique EAC-CPF records—including authorized and alternate name headings within the record, when matches are found.
- use these identity records to build a publicly-accessible prototype interface, providing researchers with integrated access to archival resources and, at the same time, access to information about the creators of those archival resources and their socio-historical contexts.

³⁵⁴ ‘[T]he Cooperative may offer novel efficiencies in the description of archival resources, through the sharing of data and new opportunities to enhance access to and understanding of primary resources’ (Pitti et al. 2015, 95).

³⁵⁵ ‘We created a prototype interface for the records that linked them at their connecting points—thereby simulating the social network of the persons, families, and corporate bodies represented—and pointed to the historical resources they created and within which they were mentioned’ (Pitti et al. 2015, 78).

- provide archives, libraries and researchers with the ability to maintain those identity records; and
- make available the software developed for extracting, matching, and merging the identity records, based on the approaches developed during the project. (Pitti et al. 2015, 81)

SNAC has three related purposes: (1) for archivists, to create new and amend existing metadata records directly in the SNAC database, using the EAD CFP schema (Figure 5.1); (2) as an internal GLAMS archival research tool; and (3) as a finding aid and research tool for researchers (who cannot amend or even suggest amendments to any SNAC records they find to be inaccurate). The Cooperative attaches persistent identifiers to each SNAC record,³⁵⁶ in effect creating a SNAC Archival Authority Index (AAI-UID). The SNAC Cooperative reports that building the SNAC AAI-UID equivalent was difficult due to the common problems in working with and combining person names.³⁵⁷

³⁵⁶ '[T]he system provides a persistent identifier for each of the merged records so that they can be durably referenced through the SNAC access system. The persistent identifiers are based on the Archival Resource Key specification, maintained by the CDL' (Pitti et al. 2015, 85).

³⁵⁷ 'An additional challenge is presented by relative quantities of descriptive data found in each record. Some records have as many as fifty or more alternative names, scores of subject headings, more than fifty related persons, families, or corporate bodies, and many linked archival finding aids or titles. Other descriptions are quite brief, based on the name occurring in one finding aid and failing to match an authority record. Finding the right method for displaying and facilitating navigation of this data presents many challenges' (Pitti et al. 2015, 89).

What is (behind) SNAC?

A schema known as **EAC-CPF** (Encoded Archival Context for **Corporate body, Person and Family** names) that created or are documented in historical resources (primary source documents), and their connections to one another in primary source documents.

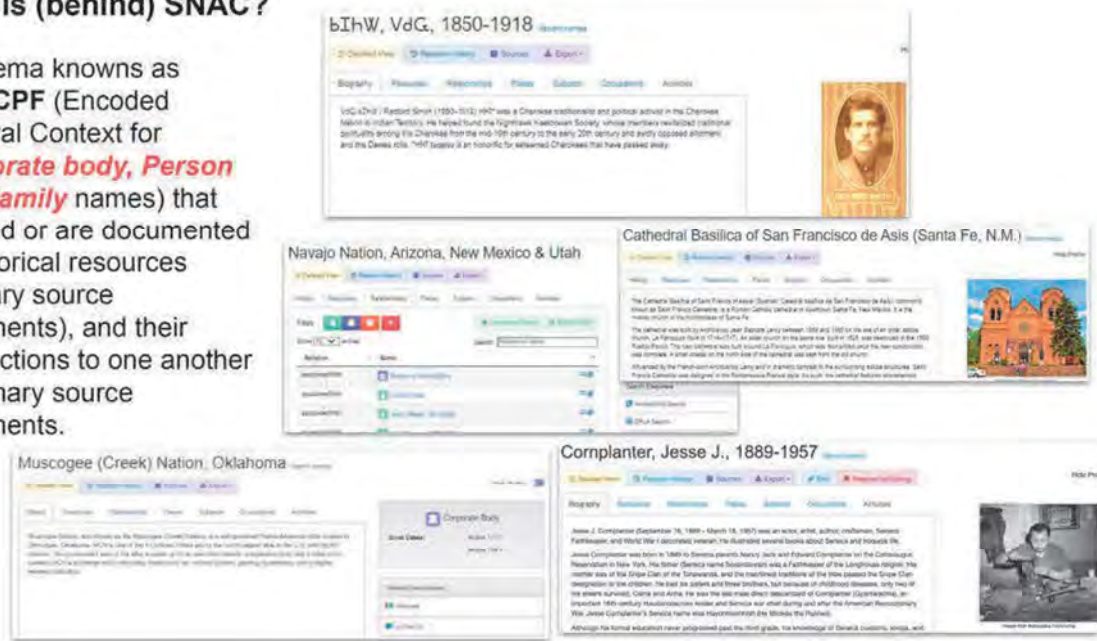


Figure 5.1 SNAC is based on the EAC-CPF Schema (<https://portal.snaccooperative.org/node/371>, accessed 19 August 2024)

This thesis does not consider the effectiveness of SNAC as a utility for GLAMS staff – it only considers its usefulness to researchers. However, SNAC offers the same research tool to both archivists and researchers (Figure 5.2).

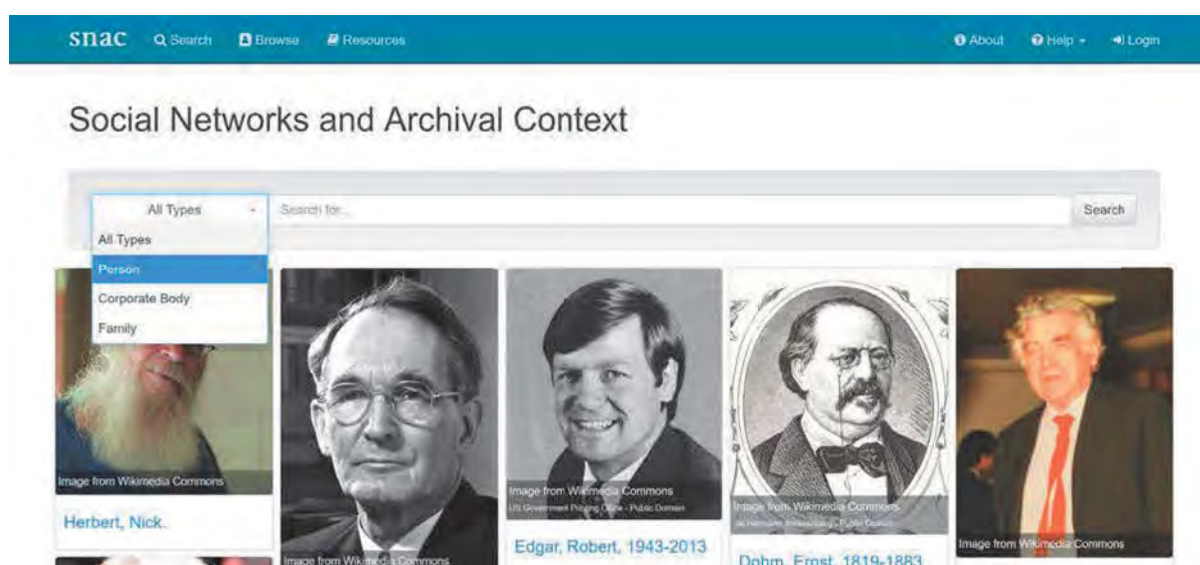


Figure 5.2 SNAC History Research Tool (<https://snac-web.iath.virginia.edu>, accessed 10 August 2024)³⁵⁸

The tool performs searches of the SNAC combined database for instances of an individual person name, a family name or a corporate body. Enquiries permitted are simple ‘search only’ or ‘expanded searches’ – similar to those found in digital archival catalogue search engines – which can be used to refine search results or limit the number of results returned by the enquiry (see Figure 5.3).

³⁵⁸ See also <https://portal.snaccooperative.org/node/332> (Accessed 10 August 2024): ‘The History Research Tool (HRT) is an aggregate of biographical information about people, both individuals and groups, who created or are documented in historical resources. Users can search for names of individual people, organizations, and families; browse featured descriptions; and discover and locate connected historical resources. Search results can be filtered by occupation and subject. The HRT foregrounds descriptions of the makers of history: persons, organizations, and families, and links these descriptions to the distributed historical resources that document their lives and work. The descriptions are biographical, and through interrelations with one another, reveal the social, professional, and intellectual networks within which the lives were lived. By linking the descriptions of the persons, organizations, and families to descriptions of archival holdings distributed around the world, the HRT provides researchers with a tool that conveniently integrates access to primary resources. With over 3.7 million descriptions, the HRT has achieved sufficient scale to be both a useful reference source, and a means to locate millions of historical resources located in more than 4,000 repositories around the world.’

Hodgkin, Thomas, 1798-1866 [Variant names](#)

[Detailed View](#) [Revision History](#) [Sources](#) [Export](#)

[Biography](#) [Resources](#) [Relationships](#) [Places](#) [Subjects](#) [Occupations](#) [Activities](#)

English physician and historian.

From the description of Papers, 1850 and undated. (Duke University). WorldCat record id: 35091897

Thomas Hodgkin was born in London in 1798, the son of John Hodgkin (1766-1845), a private tutor. The family were strong Quakers and originated in Warwickshire. He trained in medicine at Edinburgh University, taking his M.D. in 1823. After travels in Europe he became Curator of the Medical Museum and Inspector of the Dead at Guy's Hospital, London. His pathological work led him to the first description of what is now known as Hodgkin's Disease in his honour. He left Guy's Hospital following his failure, in 1837, to be appointed Assistant Physician and after a short period at St. Thomas's Hospital devoted himself to private practice and to his other interests. He had a keen interest in the world beyond Europe and in particular in the societies there that were threatened with cultural extinction by the spread of European commercial, political or cultural dominion; his works in this area included playing a moving role in the foundation and functioning of the Aborigines Protection Society. In 1850 he married Sarah Frances Scaife, a widow, from Nottingham. The couple had no children of their own but there were two sons from her first marriage. He died in 1866 at Jaffa while on a journey with his friend Sir Moses Montefiore (1784-1885) to negotiate for better treatment for Jewish residents in Palestine.

From the guide to the Papers of: Hodgkin, Thomas (1798-1866), 1840-1979. (Wellcome Library)

Figure 5.4 SNAC main return for person enquiry Thomas Hodgkin (1798–1866)

(<https://snaccooperative.org/view/61687138>, accessed 07 August 2024)

The SNAC record for Thomas Hodgkin includes:

- A Detailed View of control data.
- Revision History, a record revision log.
- Sources of the data content (including WorldCat, VIAF, Harvard.edu, ArchivesHub.ac.uk, American Philosophical Society).
- Export (JSON or EAD – CPF XML).
- Variant names.
- Associated places (Israel and Palestine).
- Subjects – a taxonomic list of academic subject tags.

- A brief biography, taken from the guide to the Papers of: Hodgkin, Thomas (1798–1866), 1840–1979 (Wellcome Institute Library, London).
- Links to other datasets and URIs for the record (Figure 5.5).
- As at the time of writing, Occupations and Activities are not functioning.



Figure 5.5 SNAC Person Record for Thomas Hodgkin (1798–1866)³⁵⁹

The sources for the SNAC record are listed in Figure 5.6. Although fully referenced on the results main page, the source is not listed under the Sources tab.

³⁵⁹ Additional information including ARK and SNAC IDs and helpful links available at <https://snaccooperative.org/view/61687138> (Accessed 07 August 2024).

Role	Title	Holding Repository	
referencedIn	Bingham, W. mss., 1752-1891	Lilly Library (Indiana University, Bloomington)	i
creatorOf	Hodgkin, Thomas, 1798-1866. Letter. 20-11-1843, [London, Eng.] to Tho[ma]s J. Pettigrew.	Haverford College Library	i
creatorOf	Hodgkin, Thomas, 1798-1866. Papers, 1850 and undated.	Duke University, Medical Center Library & Archives	i
referencedIn	John Edward Gray papers, 1783-1884, 1783-1884	American Philosophical Society	i
referencedIn	Kass, Edward H. (Edward Harold), 1917-. Papers, 1908-1990.	Harvard University, Medical School, Countway Library	i
referencedIn	Morrison, H. (Hyman), b. 1881. Hyman Morrison papers, 1899-1970 (inclusive) 1920-1963 (bulk).	Harvard University, Medical School, Countway Library	i
referencedIn	Morton, Samuel George, 1799-1851. Papers, 1819-1850.	American Philosophical Society Library	i
creatorOf	Papers of: Hodgkin, Thomas (1798-1866), 1840-1979	Wellcome Library	i
referencedIn	Papers, 1908-1990.	Francis A. Countway Library of Medicine	i
referencedIn	Samuel George Morton Papers, 1819-1850	American Philosophical Society	i

Showing 1 to 10 of 10 entries

Previous 1 Next

Figure 5.6 SNAC data sources for person name enquiry Thomas Hodgkin (1798–1866)

(<https://snaccooperative.org/view/61687138>, accessed 07 August 2024)

SNAC also offers network visualisation technology to display the related data returned from the enquiry as a network graph, probably using FOAF layout³⁶⁰ (because the graph displays with a ‘degrees of separation’ filter common to the FOAF ontology; see Figure 5.7).

³⁶⁰ The SNAC website is thin on detailed explanations, for instance there is no explanation of the chosen data visualisation typology or embedded settings.

Significant problems easily arise with the FOAF methodology because it lacks a robust definition of relationships.³⁶¹ If the dataset searched is unbounded by the enquirer (looking instead across the entire database), then as degrees of separation increase, the number of results quickly becomes unmanageable. Refining the search helps, and this can be done after the initial search if it returns too many results to handle visually. However, such crude refinements can easily result in unknowingly losing important data finds, simply to achieve a manageable set of results. SNAC does not allow complex user-defined queries to help address this shortcoming, for instance by allowing the researcher to ‘hold on’ to some items when data resolution simplification options are selected. Neither does the system identify different types of metadata (for example by colour coding items) so that the researcher can understand the weightings or data privileges implicit in the embedded visualisation technology. This becomes a major concern for researchers because the database consists of assimilated metadata from many datasets, resulting in the presentation to an enquiring researcher of a ‘black box’ distillation of archivist records of variable quality.

³⁶¹ ‘Not only is there no specific, unified theory of social networks; there is no specific object of social network analysis. There is a fundamental ambiguity here, as the network analysis of sociologists is certainly born out of questions about social relationships and “sociometric” studies of, e.g., friendship in classrooms. Today, however, studying social relationships and using formal network analysis are too different, only partially overlapping tasks. There are many ways to study social relationships or “social capital”, from the close scrutiny of love letters to regressions on the number of associations at the country scale; only those that are interested in the precise pattern created by one or a few sorts of ties between a set of individuals will profitably use formal network analysis. Conversely, it is sometimes interesting to describe in a relational way things that we would not spontaneously describe as social relationships, such as the routes of ships between ports or the fact of sharing the use of key words’ (Lemerrier 2015, No page numbers).

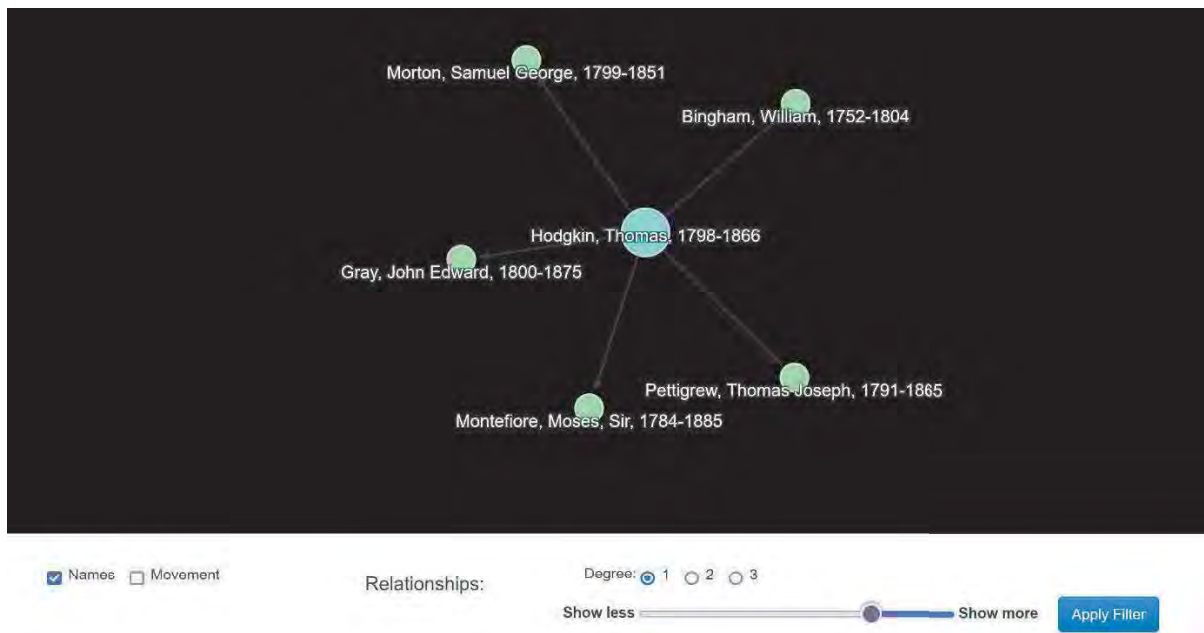


Figure 5.7 SNAC – enquiry into the name ‘Thomas Hodgkin’, degree 1
https://snaccooperative.org/visualize/connection_graph/61687138/9024591, accessed 7 August 2024)

Figure 5.7 shows the SNAC visualisation of the results of an enquiry for the name ‘Thomas Hodgkin MD’. The graph can be widened out by size (labelled degrees 1, 2 and 3). In Figure 5.7 the smallest dataset of one degree of separation is displayed. Figures 5.8 and 5.9 show a display of the same data with two and three degrees of separation. The SNAC website and related supporting materials do not explain the formula for the division of data in the visualisations into degrees, but given that the URL does not change when different degree selections are made, this suggests that the division is based on the visualisation technology degree of separation choice alone, rather than displaying (say) different categories of relationship.

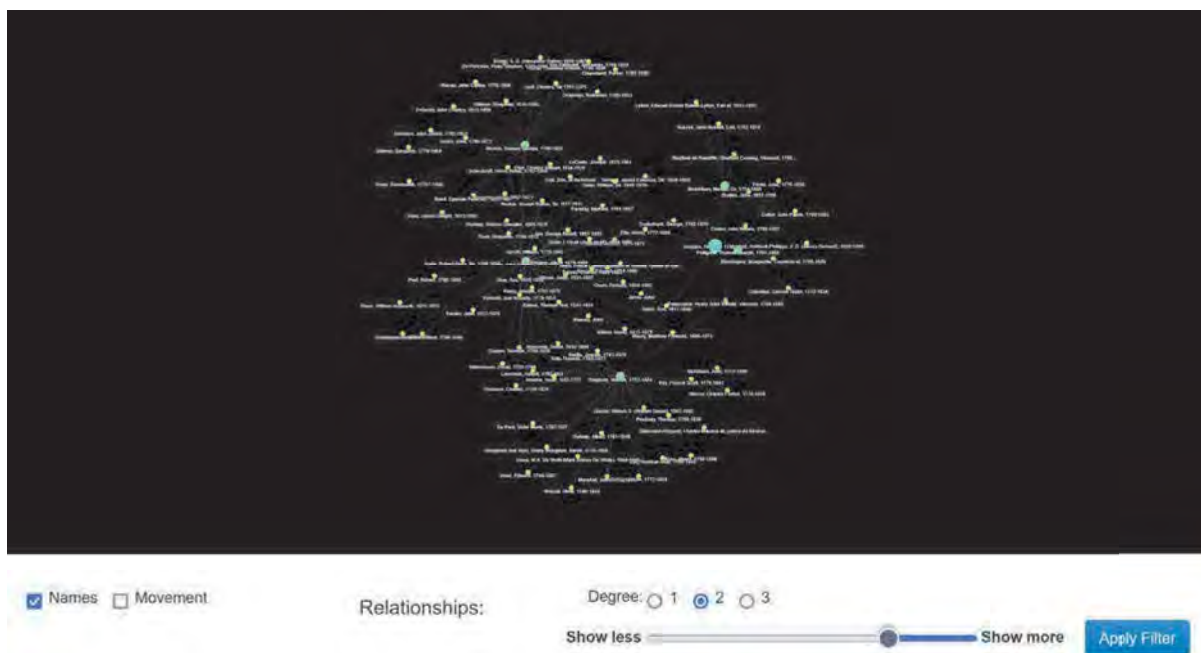


Figure 5.8 SNAC – enquiry into the name ‘Thomas Hodgkin MD’, degree 2
https://snaccooperative.org/visualize/connection_graph/61687138/9024591, accessed 7 August 2024)

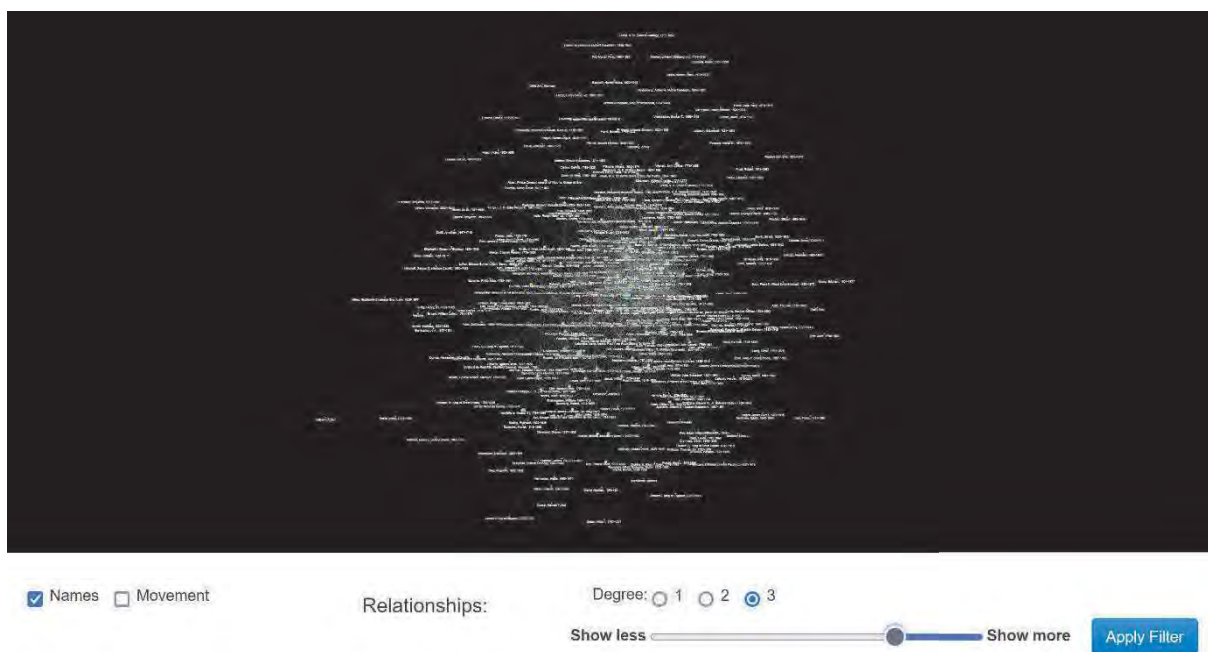


Figure 5.9 SNAC – enquiry into the name ‘Thomas Hodgkin’, degree 3
https://snaccooperative.org/visualize/connection_graph/61687138/9024591, accessed 7 August 2024)

A detailed examination of the search results for Thomas Hodgkin (1798–1866) illustrates the utility of SNAC and also its limitations and flaws, when used as a tool by a general researcher.

Thomas Joseph Pettigrew (anatomist and Egyptologist)³⁶² is clearly visible in the graph at degree 1 level (both as a friend and a friend of a friend). Examining the data generating this relationship plot reveals that both he and Hodgkin wrote two separate and unconnected letters, at different times, each to the zoologist John Edward Grey, and this is the basis of their relationship in SNAC. Pettigrew seems to appear in the graph solely because like Hodgkin he once wrote to Grey, not because he has a direct relationship of any sort with Hodgkin (he might have one via his career as an anatomist, but the system does not show it). From the data returned by the enquiry it is quite likely that Pettigrew did not personally know Hodgkin, so his appearance in the graph is simply because he is a ‘friend of a friend’, though ‘friend’ is a questionable attribute for two people who merely wrote on separate occasions and on unconnected matters to the same third party. Pettigrew’s appearance in the graph could be considered by many researchers as ‘noise’.

Hodgkin’s connection to William Bingham is not clear. According to the graph and supporting data, Hodgkin was 5 years old when Bingham died in 1804. The link to the original source from Indiana University gives dates for the Bingham papers as 1752–1891.³⁶³ Indiana University’s Bingham record allocation must contain a related persons access point

³⁶² James Marshall and Marie-Louise Osborn Collection, Beinecke Rare Book and Manuscript Library, Yale University.

³⁶³ Bingham, W. mss., 1752–1891. Papers of William Bingham at the Lilly Library, Indiana University, Bloomington, Indiana. https://webapp1.dlib.indiana.edu/findingaids/view?doc.view=entire_text&docId=InU-Li-VAA1948 (Accessed 10 August 2024).

link to Thomas Hodgkin MD (1798–1866) and this link should be questioned. Close re-examination of the source document by this researcher suggests that this correspondence item is most likely between Bingham and Thomas Hodgkin Snr (1741–1815), Thomas Hodgkin MD’s grandfather.

Samuel George Morton of Philadelphia was a Quaker, and he had shared scientific interests with both Hodgkin and James Cowles Pritchard (Hodgkin’s scientific mentor). Hodgkin corresponded with Morton with an aim to bring American native aborigines to England, in line with Hodgkin’s known and provable interests. Hodgkin also had a deep, long-lasting and provable relationship with Moses Montefiore.³⁶⁴ Only Morton and Montefiore in this graph have meaningful relationships with Thomas Hodgkin MD. One other relationship is most likely an error, and the other two are FOAF relationships of questionable research value.³⁶⁵ Therefore, the SNAC graph does not provide a representative picture of the life and achievements of Thomas Hodgkin MD at the FOAF degree 1 level.

Degree levels 2 (Figure 5.8) and 3 (Figure 5.9) add a lot more relationships and it would be prohibitively difficult, but absolutely necessary given the analysis above, to verify them all. It is doubtful whether SNAC could reveal much useful information about the life of Thomas Hodgkin, especially because all SNAC results would have to be confirmed by direct reference to each of the sources.

It is not visually possible in SNAC to easily identify or mark in the model Thomas Hodgkin’s two credible prosopographical relationships or to separate prosopographical relationships from the wide range of inferential and speculative relationships that are in SNAC mixed in

³⁶⁴ <https://hekint.org/2020/10/20/the-quaker-and-the-jew-an-enduring-and-impactful-friendship-thomas-hodgkin-and-moses-montefiore/> (Accessed 10 August 2024).

³⁶⁵ Chapter 4 has shown that FOAF ontology is valuable for Born Digital data but of less utility in the research of historical data. It is arguably a poor choice of graph layout for SNAC.

with EBP relationships. In SNAC, all archival records, irrespective of research value, have equal weighting.³⁶⁶

Furthermore, an alternative desktop search, using common search engines and online finding aids, for data on the life of Thomas Hodgkin quickly establishes that there are very many more (and more meaningful) discoverable relationships. These can be found in the Wellcome Institute Hodgkin papers (which SNAC uses only for its biography but not data) and even more at the Bodleian Libraries (several hundred relationships over several thousand items, as the P7 Case Studies show). However, these will not appear in any digital finding aid based on GLAMS records, because the vast majority of them have not been rendered into individual MARC metadata records. Their partial online accessibility is through HTML only.

Had SNAC referenced its person names to NAI-UID indexes then the confusion over similar names (here grandfather and grandson) would have been easier to discover and resolve.

The Quaker and occupational connections between Hodgkin and Morton would have emerged through the incorporation of prosopographical attribute data which could have been accessed through the NAI-UID system, and also the spurious relationships would have more easily been identified and disregarded.

³⁶⁶ 'The lack of information about the provenance of collections, or individual items, is exacerbated in digital archives and collections, or collections of digital historical representations. As Joshua Sternfeld has highlighted, items that become part of digital collections can easily get detached from their original collection context, and in that process, existing information about the original provenance of the item frequently gets lost. This can also happen with digital collections that are removed from their original creation context. Just as in many physical archives, the contextual information about the provenance of digital collections, or digital objects that are part of digital collections, may not have been collected in the first place. Supplying information about provenance in digital archives is also more complicated due to the massive scale of many collections, and due to the fact that one has to distinguish between the provenance of the original record, item, or collection (if it was a physical object that has been digitized), and the provenance of the digital historical representation, or collection of digital historical representations' (Hering 2014, 2).

Comparing SNAC to the University of Birmingham’s finding aid, which is an implementation of Exlibris,³⁶⁷ the same enquiry returned there 303 records (see Figure 5.10). This finding aid gives much more comprehensive coverage of Hodgkin’s remarkably active life as a Quaker, pathologist, ethnologist and political activist championing the rights of aborigines. Of course, this finding aid cannot find the considerable online prosopographical data on the life of Thomas Hodgkin that can be accessed on the Wellcome Institute and Bodleian Libraries websites and accessed through their finding aids, but it is still a more valuable enquiry tool for the researcher than SNAC.

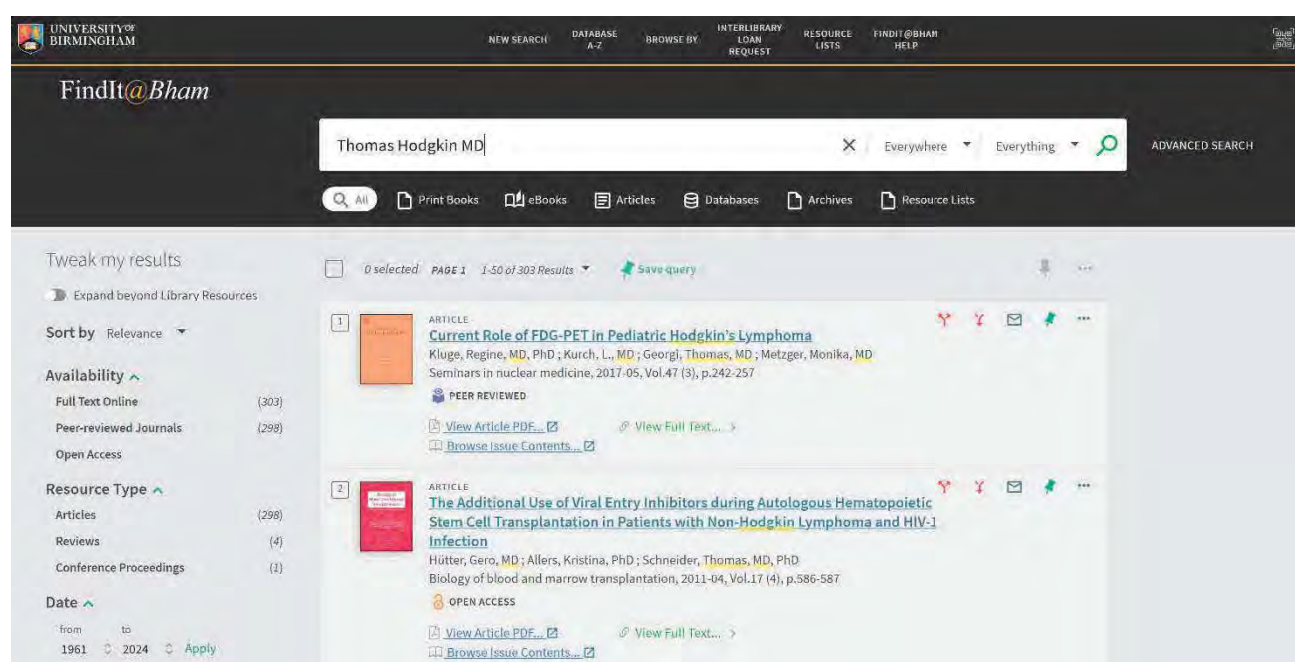


Figure 5.10 University of Birmingham (a hosting of Exlibris) finding aid search: Thomas Hodgkin MD (https://birmingham-primo.hosted.exlibrisgroup.com/primo-explore/search?query=any,contains,Thomas%20Hodgkin%20MD&tab=local&search_scope=CSCOP_44BIR_DEE_P&vid=44BIR_VU1&offset=0, accessed 10 August 2024).

³⁶⁷ Exlibris is a part of the Clarivate/Proquest/Web of Science organisation. <https://exlibrisgroup.com/products/primo-discovery-service/content-index/> (Accessed 10 August 2024).

It must be recognised that SNAC, which is a relatively new affordance in DH, will take time to develop, and that development is conditional on both the location and interest of participating archives, and on improving the LOD connections made to other archives and other finding services.³⁶⁸ The current incompleteness of SNAC is fully recognised by the Cooperative (it is fair to describe it as a work in progress). To balance the critical analysis above, SNAC's recognition of its limitations is reproduced here in full:

Although SNAC demonstrates the potential professional and scholarly benefits of the novel research tool, it is clear that computer-based techniques alone cannot fully realize the transformative potential that this resource tool offers researchers. Archivists and librarians did not create EAD finding aids with SNAC's use of the data in mind. Thus the quality of the data is uneven and many potential interrelations of individuals, families, corporate bodies, and associated historical documents are unavoidably overlooked. The quality and thoroughness of the research tool demonstrated in SNAC, while also compelling, reflects its data sources and is uneven and incomplete. (Pitti et al. 2015, 93-94)

The Cooperative that produced and maintains SNAC is deliberately multi-skilled, including a variety of appropriate skills in the development team.³⁶⁹ The 'Independent Researcher' approach taken in the P7 Case Studies is a similar but much smaller version of that offered by the much larger and better supported project SNAC. Unlike the SNAC offering, the P7 HTTD and the P7 Case Studies combined datasets are incomplete, but P7 is based on EBP instead of a mixture of EBP and metadata, and as such produces a more robust affordance offering to the researcher than SNAC. It will be shown in Chapter 6 that the HDDT offers a

³⁶⁸ Currently all participating members are US institutions.

³⁶⁹ 'The original internally created personas served as a reference point for assessment recruitment, and participants were ultimately derived from the following user types: researcher (genealogist, academic researcher) and reference staff (as researcher surrogate), archival/library staff (processing, description, cataloging, and administrator), and other users with interest in using the network of data amassed within SNAC for digital humanities project interests' (Pitti et al. 2015, 93).

much richer and more meaningful analysis and visualisation of data while using broadly similar approaches and technologies.

In summary, SNAC is a research affordance of proven utility to GLAMS archivists to manage shared metadata, but it is of limited utility for non-GLAMS researchers. This is because:

- The underlying data found is of variable quality.
- The visualisation tool (based on FOAF ontology) is of questionable value as a research tool because it produces 'related' finds between unrelated persons and mixes EBPD with expert opinions through metadata.
- Only a very small number of finds are made compared to the number of qualifying records actually held at Cooperative institutions.
- Researchers are unable to devise their own queries or manipulate search results.

5.3 The Cambridge Group for the History of Population and Social Structure and the I-CeM project

The Cambridge Group was formed in 1964 'to undertake quantitative research in family history and demographic history'.³⁷⁰ Since 1964 The Cambridge Group has had a major influence on population studies made using digitised census enumerators' books, supplemented with other related population survey data (see Figures 5.11 and 5.12).³⁷¹

³⁷⁰ <https://www.campop.geog.cam.ac.uk/> (Accessed 5 August 2024).

³⁷¹ 'The Group's long tradition in the analysis of nineteenth and early twentieth-century census enumerators' books (CEBs) has built up the largest pool of researchers using British CEBs anywhere. The availability of digitised versions of all the censuses from 1851 to 1911, through the ICeM project run by Kevin Schurer (Leicester) and Eddy Higgs (Essex), presents a range of opportunities for research in demographic, economic and social history.' <https://www.campop.geog.cam.ac.uk/about/history> (Accessed 5 August 2024).

[illegible]

Figure 5.11 1851 Household Schedule England and Wales (English) – reverse. Instructions for enumerators (Higgs et al. 2021, 23)

[Example of the manner in which Entries should be made in the Inmate Book.]

153

Parishes Township of		Ecclesiastical District of		Civil Parish of		Village of		Village of	
of		of		of		of		of	
Name of Person, Thence, or		Name and Residence of each Person		Residence		Rank, Profession,		When Born	
and Thence, or		as the Right of the 40th Statute, 1700		Habit of Family		or			
the of House						Occupation			
Anglican House		William Johnson		Bachel	25	Bachel, in Service		1700	
		John Doe		Wife	20			1700	
		John Doe		Son	10			1700	
		William J. Doe		Son	15	Bachel		1700	
		Thomas Doe		Son	5	Son		1700	
The House Building									
Anglican House		John J. Doe		Bachel	20	Bachel, in Service		1700	
		Mary Doe		Wife	15			1700	
		John Doe		Son	5	Son		1700	
		Thomas Doe		Bachel	10	Bachel, in Service		1700	
		Mary Doe		Wife	15			1700	
		John Doe		Son	5	Son		1700	
		Thomas Doe		Son	10	Bachel, in Service		1700	

Working with census data where paper has partially decomposed over time, and instructions to enumerators varied from census to census depending on government data

interests at the time, was very challenging.³⁷² Peter Laslett, a leading member of The Cambridge Group, and Kevin Schürer, a founder member of the I-CeM project (see Section 5.2.1), offer encouragement that the scope of EBP is now within the reach of research. However, they do not offer suggestions as to how to attempt the work itself.³⁷³ EBP as set out in this thesis shows how this opportunity might be seized.

The Cambridge Group does, however, have a comparable interest in the systematic study of the annual reports of the Registrar General for aggregate census data, 'Histpop'.³⁷⁴ This is a different dataset to that which EBP initially recommends for building the NAI-UID index, but it is a related source of comparable content and quality. Laslett recognises this lack of take-up of BMD records as an object of research, noting that it is both an important omission and an important opportunity.³⁷⁵ Laslett also recognises the desirability that in the future, historians work within (and lead) small groups of specialist researchers (this thesis calls

³⁷² 'The GRO and GROS had comparatively little time to organise the taking of the census, and some of the agents involved left much to be desired – illiterate householders, slap-dash enumerators, and registrars who did not supervise the work properly. This alerts us to the problematical nature of some of the data in the manuscript returns. The information in the enumerators' books was several stages removed from reality, and each stage could add its own accumulation of errors' (Higgs Edward et al. 2021, 16).

³⁷³ 'In order to undertake genetic analysis, it is necessary to be able to study a community of specimens over a number of generations, the more the better. The great disadvantage of human beings for genetic study is that the generation is so long. With drosophila it is possible to observe the passage of ten generations in a matter of days. With humans ten generations would take some 300 years to observe. Now 300 years happens to be within the period during which the registration of births, marriages and deaths can be studied from this evidence, and in England we can go back two or three generations further. It is difficult as yet to see how this opportunity might be used' (Laslett and Schürer 2021, 302).

³⁷⁴ 'Histpop' includes all annual reports of the Registrar-General (of England and Wales) that were published before the change of title to the Registrar-General's Statistical Review in 1921.

³⁷⁵ 'The phrase "sociological history" has been occasionally used here as its title, but it might almost be better to use "social structural history" instead. This new title is required first and foremost to register a distinction in subject matter, for confessedly historical writing has not previously concerned itself with births, marriages and deaths as such, nor has it dwelt so exclusively on the shape and development of social structure. But the outlook is novel as well as the material, at least in its emphasis. Perhaps the distinctive feature of the attitude is the frank acceptance of the truth that all historical knowledge, from one point of view, and that an important and legitimate one, is knowledge about ourselves, and the insistence on understanding by Contrast' (Laslett and Schürer 2021, 295).

them Independent Research Groups).³⁷⁶ EBP brings these two ideas together, exploiting BMD records to establish a suitable authority system in the use of person names.

The Cambridge Group has produced many datasets, including fifty-three in the Occupations series, and many of these are available online through the Online Historical Population Reports (OHPR) website (hosted by the University of Essex).³⁷⁷ Other datasets are hosted by the UK Data Service (UKDS) and are available for other researchers to reuse. The background to the work of The Cambridge Group could not be investigated via the website.³⁷⁸ Guidance, parameters and definitions for individual datasets are, however, included with data downloads.

Some of the complexities faced by The Cambridge Group in (for instance) data classification determinations are set out in one paper accessible from the website.³⁷⁹ While the data collected by The Cambridge Group is of proven value, the considerable extent of data

³⁷⁶ 'The historian cannot hope to make his contribution to studies of this sort unless he is rather differently equipped than he has previously been ... What must come into being is a working community where the historian is in the confidence of the economists, the statisticians and the others. Nevertheless the responsibility for enabling us all to understand ourselves in time must still rest where it has always rested, on the historian as an individual' (Laslett and Schürer 2021, 302).

³⁷⁷ 'Histpop – The Online Historical Population Reports Website. A collection of British Historical Population Reports. The Online Historical Population Reports (OHPR) collection provides online access to the complete British population reports for Britain and Ireland from 1801 to 1937. The collection goes far beyond the basic population reports with a wealth of textual and statistical material which provide an in-depth view of the economy, society (through births, deaths and marriages) and medicine during the nineteenth and early twentieth centuries. These 200,000 pages of census and registration material for the British Isles are supported by numerous ancillary documents from The National Archives, critical essays and transcriptions of important legislation which provide an aid to understanding the context, content and creation of the collection. In digitising this resource the OHPR has enabled Browsing through the collection by date or geography, or Searching the content directly. Documents relating to the digitization and web development process may be accessed via the Project tab. OHPR is an AHDS History project, funded as part of the JISC Digitisation programme and is hosted by the UK Data Archive at the University of Essex. Note to genealogists and others tracing individuals: This site only contains a very small number of reproductions of original census enumerators' books for illustrative purposes.' <http://www.histpop.org/ohpr/servlet/Show?page=Home> (Accessed 10 September 2024).

³⁷⁸ Unfortunately, none of the links on the Project Documents page of the website is working at the time of writing. (They return HTTP Status 404 – Not Found.)

³⁷⁹ The PST system of classifying occupations, paper by E. A. Wrigley. <https://www.campop.geog.cam.ac.uk/research/occupations/datasets/coding/> (Accessed 5 August 2024).

manipulation and necessary data systematisation, as well as the structuring of data into manageable datasets capable of integration, inevitably mask much of the original qualities of the underlying data. Accurate data representation at the individual person record level is, however, not a major concern for The Cambridge Group, because its datasets are intended to be used for statistical aggregations of person prosopographical attributes. Therefore, the qualities of fixity or representational quality of digital resources relative to the sources do not arise as a significant concern. Nevertheless, the data must be (and frequently is) reused with care. Unfortunately, The Cambridge Group is unable to offer support to future researchers at a substantive level and this is a concern for the long-term usability of the data.³⁸⁰

The difficulties are universal, and The Cambridge Group has made significant steps to overcome these, as the evidence of the popularity of its datasets indicates. In making prosopographical datasets, the linkage (and routing) to the underlying primary data requires significant management resources and considerable supporting literature. These data to source linkages can erode over time (in this case outside of the control of The Cambridge Group). The solution is for the EBP at sources (BMD or Census records) to be indexed through an NAI-UID system. These are now nearly fully represented digitally at GRO. This would allow all research data records that reference persons to be allocated an RAI UID

³⁸⁰ 'Most of the datasets listed below will be available from the Economic and Social Data Service at the UK Data Archive shortly. However, some of them may be embargoed for a period. The datasets all come with adequate documentation for those with the requisite technical knowledge. Sadly, we do not have the resources to provide further assistance. However, we recognise that some scholars may wish to use these datasets, who either lack the requisite technical skills, or require further assistance for other reasons relating to the complexity of these resources . We are investigating mechanisms by which we might be able to fund such assistance and we would encourage registering expressions of interest in case such funding does become available (and registering an expression of interest might help us to secure funding in due course). In the meantime, we may be able to help if you have funding available or if the query requires very little of our time.' <https://www.campop.geog.cam.ac.uk/research/occupations/datasets/catalogues/occupationspopulation> (Accessed 5 August 2024).

linked to the NAI-UID. Research project data management and classification allocations could then be recorded at the individual person level in The Cambridge Group's datasets, providing a reliable route to the original GRO record.

The classificatory actions taken by The Cambridge Group in forming the datasets are, from a new researcher's perspective, hidden in the background, with the considerable machine and human work undertaken hard to assess without careful study. If, for instance, a researcher took a subset of data from one of The Cambridge Group's datasets and then combined it with another subset of data from another research group's dataset, then the new researcher would in effect be creating a new secondary dataset of mixed origin. The difficulty of making classification determinations for the new dataset, in such a way as to fully embrace the classifications made by (in this case) The Cambridge Group, and combining them with those of another data source could be both complex and hard to explain, with potentially three different data management schemes in hand (the Cambridge Group's, the other contributing dataset and the combined dataset).

5.3.1 The Integrated Census Microdata project

This section looks in detail at the I-CeM project,³⁸¹ a collaborative venture between the University of Essex and FindMyPast, a commercial provider of a genealogical website. The

³⁸¹ 'The 1-CeM project was path breaking in bringing together commercially produced data created essentially for a fee-paying genealogical and family history audience, rendering and repurposing them for academic based researchers, teachers and learners ... The National Archives (TNA) in London launched its own website for the 1901 census in 2002 but later took the decision to enter into licensing arrangements with FindMyPast to make the 1911 census for England and Wales available online. This public-private partnership also proved to be critical for the future I-CeM project since TNA worked with Schurer, who was then Director of the UK Data Archive (UKDA) at the University of Essex, to help broker access to the commercial data created by FindMyPast in order to create a fully enhanced integrated research and teaching resource for use by the academic community' (Higgs et al. 2021, 145).

venture is an exemplar for academic/commercial cooperation and it is directly relevant to this study of EBP, because it uses comparable national data, in part sourced this time from a copyright FindMyPast genealogical dataset (names and addresses) initially developed by TNA.³⁸² The project was funded by the UKDS,³⁸³ with the first version of the dataset published in 2014. The project team produced a guide which ‘has three main purposes’:

- first, to explain the history of census taking in Great Britain, the documentation used in that activity over time, and the official publications produced;
- secondly, to explain the provenance and construction of the 1-CeM data collection;
- lastly, to describe the structure of, and access to, the 1-CeM data collection. (Higgs et al. 2021, 1)

Each of the censuses from 1851 to 1911 was highly structured and undertaken by trained census takers who were supported by detailed instructions. Nevertheless, the research team commonly found messy data.³⁸⁴ There had also been significant data loss over the time period of the study, although much ‘lost’ data has subsequently been re-found.³⁸⁵ As a

³⁸² ‘The I-CeM data collection was generated to support research across a number of disciplines across the humanities and social sciences, and designed to facilitate research over time, by region, and nationally, including comparative analysis alongside international research resources, where they exist. The Integrated Census Microdata (I-CeM) project was based within the Department of History at the University of Essex and funded by the Economic and Social Research Council (ESRC RES-062-23-1629) between 2009 and 2013. It created an integrated collection of census microdata with 100% coverage of the decennial censuses of England and Wales for the period 1851 to 1911, and for Scotland for the period 1851 to 1901 using data generously supplied to the project by the genealogical internet provider FindMyPast’ (Higgs et al. 2021, i).

³⁸³ <https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=7481#!/access-data> (accessed 25 October 2023).

³⁸⁴ ‘The GRO and GROS had comparatively little time to organise the taking of the census, and some of the agents involved left much to be desired – illiterate householders, slap-dash enumerators, and registrars who did not supervise the work properly. This alerts us to the problematical nature of some of the data in the manuscript returns. The information in the enumerators’ books was several stages removed from reality, and each stage could add its own accumulation of errors. The household schedules that form the 1911 returns in England and Wales may be closer to “raw” data but might still contain inaccuracies. However, despite their imperfections, the census returns, and the I-CeM data collection based on them, are still a wonderful source for historians. (Higgs et al. 2021, 16).

³⁸⁵ ‘There have inevitably been some losses amongst the original returns, which have not always been held in optimum conditions. The backs and fronts of some of the enumerators books have been damaged by storage on unsuitable racking, and in some cases there has been more serious damage and loss. But some of these gaps are slowly being made good, as in the case of the 1851 census returns for Manchester, Salford, Oldham

result, and to make the data outputs usable for research purposes, data underwent a significant and rigorous cleaning process,³⁸⁶ and ‘this process had the following main functions’:

- to reconcile the data with the Census Reports;
- to reformat the input data;
- to perform a number of consistency checks on the data, and to alter the results accordingly;
- to reformat and standardise the data;
- to add a number of enriched variables, mainly relating to household structure. (Higgs Edward et al. 2021, 146)

The project used computing with algorithms to perform data cleaning exercises in stages, because ‘the entire I-CeM database for Great Britain, runs to prosopographical data on some 210 million person and includes forty-five million household records’.³⁸⁷ The Handbook for the project includes a detailed explanation of the careful and comprehensive

and Ashton-under-Lyne, which were severely damaged by flooding. These have been patiently transcribed by Manchester and Lancashire Family History Society, and are now available on line. These new transcribes are incorporated into the 1-CeM data collection’ (Higgs et al. 2021, 18).

³⁸⁶ ‘However, this bold description gives little insight into the accumulated work and effort that has gone into the creation of the final dataset. The multiple computer programs used to manipulate the data included thousands of lines of code, drawing on many years of experience by Schurer in the computer-based analysis of individual and household level census material. Equally, embedded within these programs are multiple calls to various coding dictionaries or look-up tables ..., a number of which took several person years to generate’ (Higgs et al. 2021, 146).

³⁸⁷ ‘[T]he Group began work on building a fully automatic record-linking program in which all the decisions, including the difficult cases, were to be made by the computer in accordance with a set of basic principles. Two main reasons persuaded us it was worthwhile expending time and effort to find a way to hand over to a machine a matter that at first sight might seem best left to the judgment of the historian. The first was theoretical. If the judgments we make about specific links have any claim to intellectual respectability, we ought to be able to specify the principles on which they are based. If we can do that, we can express those principles in the form of a computer program and get the machine to implement them more consistently than we can ourselves. The second reason is practical ... reconstituting the registers ... involves so many records that can be linked in so many ways that the size and complexity of the task can easily exceed the capacities even of a team of historians’ (Higgs et al. 2021, 147).

data cleaning process tasks undertaken,³⁸⁸ noting that, regrettably, compromises had to be made in data management, with some data digitally manipulated into a standard form.³⁸⁹

To deal with this level of uncertainty in data cleaning, the project took the decision to record and retain the original form of 'cleaned' data,³⁹⁰ so that researchers could, if in doubt, view an image of the original physical 'source'.³⁹¹

The I-CeM Handbook includes many images taken from microfilm copies of census enumerators' copy books held at The National Archives, London, the General Register Office

³⁸⁸ 'First, the "raw" textual strings for all of the major fields within the database were maintained in their original form. Second, new coded or classificatory variables derived from the original textual strings were then added to the database. These supplemented the original entries rather than replaced them. Third, all contextual alterations to the data undertaken as a result of the checking procedures were carried out on the coded variables rather than the original data strings. In addition, and most importantly, what are termed "inference" variables were assigned for major fields within the database. By default these were allocated a value of 0 (zero) but were given a different score if for any reason the value of the coded variable was changed as a result of the checking process. Thus, should a researcher wish to check the actions that have been taken by the enrichment program, and if necessary return to the original entries of the underlying source, they can do so via reference to the appropriate inference codes. Equally, should any researcher wish to develop their own classificatory scheme for, say, occupations or relationship to head of household, they can easily devise their own coding look up table and apply it to the respective original textual string. This general approach is good database management practice, and maintains maximum flexibility in the secondary use of the historical source material' (Higgs et al. 2021, 148).

³⁸⁹ 'Due to variance in the ways in which the same information is recorded textually, in their raw form these strings are almost impossible to analyse comprehensively unless some form of standardisation is undertaken. In tackling census transcriptions covering the entire country, this trouble becomes immense' (Higgs Edward et al. 2021, 150).

³⁹⁰ 'Finally, it is important to realise that whilst every effort has been made to ensure consistency across all the standardisation undertaken in this project, the coding is not and cannot be 100 per cent accurate. Mistakes will undoubtedly have been made. In part this is due to the fact that by its very nature, coding is a subjective exercise. Decisions over how an ambiguous string should be classified will vary from person to person. In addition, for straight forward practical reasons, all strings had to be coded "blind". That is to say that, that the strings were coded as strings, in the absence of any contextual information about the individual taken from the rest of their census record. Whilst this may not be important in the majority of fields, or cases, it may have significance in the case of occupations and relationships. With the former, because any one dictionary entry can only have one code, an occupation title which could have more than one meaning will only default to a single code. In the case of relationships, a simple string such as "son" might be nuanced by the familial situation in which it is recorded, perhaps in reality being a step-son or a grandson if there is an intervening generation. This problem, however, was addressed by a series of programs which classified households by taking all individuals within the household into consideration and re-assigning relationship designations if appropriate' (Higgs et al. 2021, 154).

³⁹¹ 'Because of this ambiguity around standardisation and coding, it is important to realise that during this process when codes were added or data is "altered", the original strings from which the codes were derived are still preserved as separate variables within the database, this users can recode or reclassify should they wish to do so' (Higgs Edward et al. 2021, 154).

for Scotland, or in some cases local record offices. It was important to I-CeM that the provenance of Representative Data be established and preserved. The I-CeM project datasets now reside with the UKDS and are freely accessible in redacted form, but with person names and addresses are viewable only under Special Licence (from FindMyPast). These are issued on a case-by-case basis, because the project committed not to link person data across censuses in order to protect the commercial property of FindMyPast. Nonetheless, the open-source dataset is of great value as a research resource. A Freedom of Information Request shows that the database has been frequently accessed and used (see Tables 5.1 and 5.2).

I-CeM downloads

Year	Number of downloads
2020	1752
2021	1003
2022	1306
2023 (up to and including 29/10)	1323

Table 5.1 Number of I-CeM downloads. Data supplied by the I-CeM team in response to a Freedom of information request by the author

Special Licence SN 7856 awards

Year	Number of licence awards
2020	11
2021	15
2022	20
2023 (up to and including 29/10)	21

Table 5.2 Number of I-CeM Special Licences. Data supplied by the I-CeM team in response to a Freedom of information request by the author³⁹²

The I-CeM project in the UK uses EBPd on a national scale (e.g. census data enumerator books) and provides a data service of proven data quality and provenance. That the datasets are popular with other researchers suggests that the NAI-UID system might likewise prove popular because it is similar in form and content. I-CeM does not reveal the UIs allocated to the incidence of every person name in the dataset. Had it done so, and had it also referenced an NAI-UID index, then the data provision would be strengthened considerably. If researchers using I-CeM data in their research projects referenced both the NAI-UID and I-CeM AAI-UIs, subsequently adding further UIs to each researcher dataset, then the EBP system would emerge.

³⁹² Special Licences are awarded by FindMyPast to users wanting access to genealogical components of the dataset, these are not available to the general public.

5.4 Traces through Time

The National Archives Traces through Time (TTT) project 2016, funded by the Arts and Humanities Research Council, initially focused on building an application to find linked records (linked by person name) across many TNA databases. The project was complex and involved bringing together many skills distributed across several archives. This presents a challenge when, after project completion, the skill contributors disperse and non-contributors are left to manage the system.³⁹³

TTT began with the datasets relating to persons enlisted for service in the Great War (1914–1918).³⁹⁴ That initial exercise comprised over half a million links based on person names and other limited prosopographical data (such as dates and locations). The project used machine learning through a set of algorithms to find name matches across the datasets,³⁹⁵ using

³⁹³ 'Computational archival science projects, such as Traces through Time, typically require a wide range of skills and experience that are unlikely to be found within a single heritage institution. The project required expertise in data modelling, statistical modelling, data mining and natural language processing as well as archival science, software engineering and user-experience design plus, of course, an understanding of the needs of researchers. Our approach was to build a research consortium drawing together relevant expertise from academia and the archives sector. Over a two-year period, more than twenty individuals from six institutions have contributed their particular skills, experience and insight to the project' (Ranade 2016).

³⁹⁴ 'The identification of a link between two occurrences of an individual in the historical record is achieved through assessing the similarity between the individual attributes of the two entities to be compared. During this project, we have worked extensively with data from World War One service records from The National Archives collections. The datasets in question were initially created by indexing the original paper documents, and our analysis is limited to those data attributes which were consistently captured by previous digitisation and transcription projects. For WW 1 data we are generally restricted to linking records based only on names and either age or date of birth. Other attributes such as place of birth and service number are sometimes available but are not consistently captured across datasets' (Bell and Ranade 2015, 24).

³⁹⁵ 'We describe an approach to identifying and incorporating common differences in textual information arising from factors such as: handwriting recognition errors, typographical errors and phonetic errors made when names are recorded. A different approach is described for dates of birth, where the algorithm must accommodate inaccuracies in recording such as mis-representation of age or rounding of declared ages. In this case, the age distribution observed for each dataset is fed back into the algorithm to support a statistical approach to calculating the likelihood that two occurrences of a person with different recorded dates of birth, in fact, relate to the same individual' (Bell and Ranade 2015, 24).

attribute information, dates and location matches to help verify the name matches achieved by the algorithms.³⁹⁶

Figure 5.13 shows an example of the TTT project's eventual absorption into TNA's 'Discovery' search engine. A search for the name Alfred Frederick Minall generates a report which includes under 'Other possible matches' the results of the TTT algorithmic matching process.³⁹⁷

The screenshot displays a web page from the National Archives 'Discovery' search engine. The main record is for 'Name Minall, Alfred Frederick Service Number: 1/304 RNVR Division: Sussex ...'. It includes fields for Reference (ADM 3376855), Description (Name: Minall, Alfred Frederick; Service Number: 1/304; RNVR Division: Sussex; Date of Birth: 10 April 1894), Date (1902-1919), Held by (The National Archives, Kew), Former reference in its original department (Vol No 1), Legal status (Public Record(s)), and Closure status (Open Document, Open Description). To the right, there is a section for 'Ordering and viewing options' with a price of £3.50 and a button to 'Add to basket'. Below the main record, there is a section titled 'Other possible matches' which states: 'The following records may contain information about the person described above. As the links are found by computer analysis, we cannot guarantee they are the same individual or that every record in which the person appears will be listed.' It then lists three matches, each with a date of birth of 10 April 1894 and a service number. The first match is 'Minall, A F' with service number 1/304. The second match is 'Minall, Alfred Frederick' with service number 1427938. The third match is 'Minall, Alfred Frederick' with service number 1427938. Each match has a 'Strong match' rating and a button to 'View record'. There is also a 'Help with your research' section on the right side of the page.

Figure 5.13 Traces through Time project record for Alfred Frederick Minall showing TTT links to 'other possible matches' (<https://discovery.nationalarchives.gov.uk/details/r/D7695214>, accessed 12 October 2024)

³⁹⁶ 'We have identified ways of linking names across records, with the added value of a confidence rating. For example, when we look at Alfred Minall's records we can be reasonably sure that they relate to the same individual. Making links is only half the task; we also need to calculate the statistical likelihood that they really are the same individual. This is based on a range of measures, such the similarity of name and dates, whether a name is unusual or whether we have other information such as service numbers.' <https://blog.nationalarchives.gov.uk/making-connections-tracing-people-collection> (Accessed 12 October 2024).

³⁹⁷ 'There was some initial concern from within The National Archives that this feature might be misleading to users' (Ranade 2016).

Figure 5.14 illustrates the extent of the algorithmic matching process across records which takes place in the background. Nodes identify instances of:

- Family name = Green – Minall
- First names = Red – Frederick OR F OR Alfred OR A
- Birth records = Pink
- Blue = TTT categories
- Grey = TTT containers (of records)

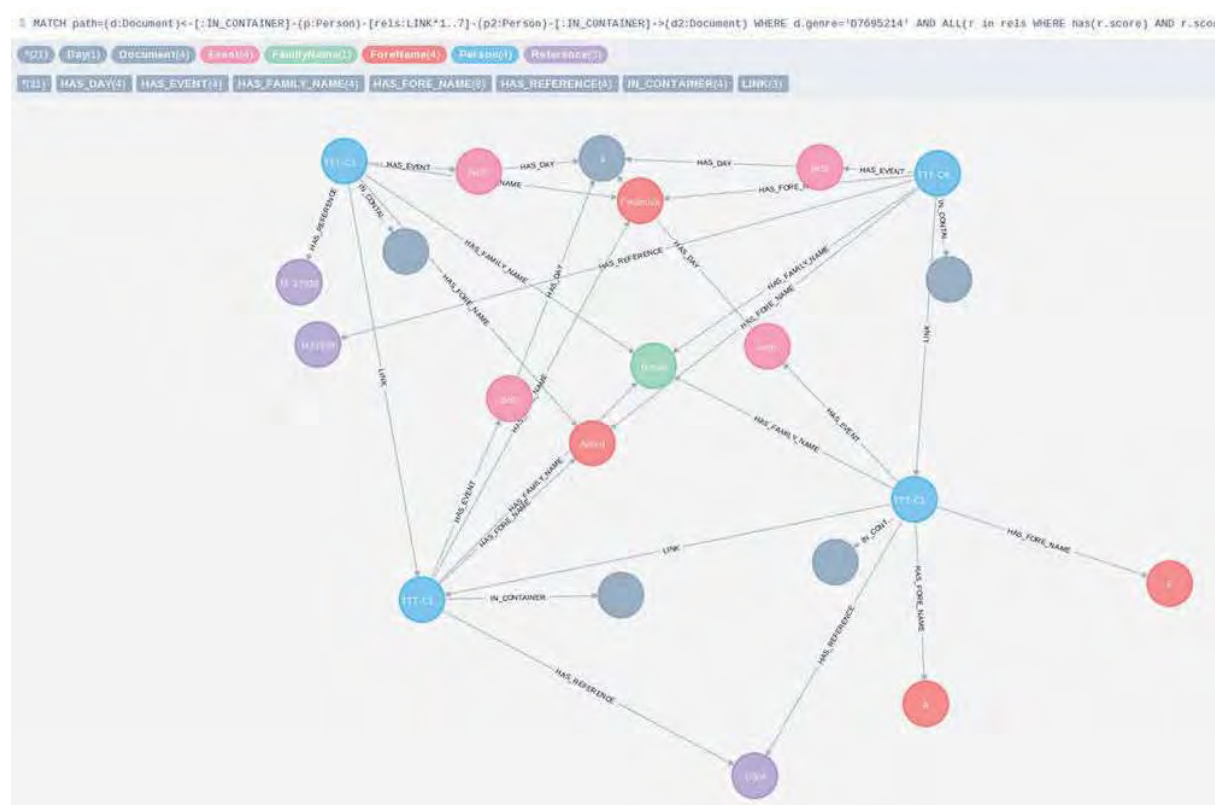


Figure 5.14 Graph showing links across records of Alfred Minall (<https://blog.nationalarchives.gov.uk/making-connections-tracing-people-collection>, accessed 12 October 2024)³⁹⁸

³⁹⁸ See also <https://blog.nationalarchives.gov.uk/making-connections-tracing-people-collection> (Accessed 12 October 2024): 'The Beta version of this new function covers twenty records series from the First World War period and over half a million newly identified are links now available through Discovery, each with an associated confidence score.'

The project used machine learning techniques, training its models on national datasets (sourced from GRO records), and sought probability calculations through a combination of name matching algorithms (spelling variations) and first and last name cultural combinations (Angus Stewart is a likely name for a person born in Scotland) to help overcome illegibility instances in the source data.³⁹⁹

The project hoped to extend this work further, from its current limitation of single person record linkages, to be able to allow users to explore complex social networks (relationships) by including related persons in the matrix.⁴⁰⁰ Discussions with TNA staff indicate that the project failed to progress because of concerns with the complexity of algorithmic solutions used to underpin data based on general public offerings, and also a concern that much historical prosopographical data is inevitably interpretative (even subjective), given the messiness of the primary data. A data match made by a TNA array of algorithms has no more validity than any other comparable matching process,⁴⁰¹ especially in matching data derived from the Records of PHL. Essentially, person name matching by algorithm produces a probability matching regime which works well over large datasets, but can only be a

³⁹⁹ 'The project described here aimed to begin this transformation by extending the boundaries of current computational archival science research in three important directions: to increase the extent and diversity of the data that can be handled using modern data-analytical techniques; to improve support for the "fuzzy" data that is typical of archival collections (i.e. data that is incomplete, inaccurate, inconsistent or uncertain); and to develop robust confidence measures for the links we identify, enabling archivists to qualify the assertions we make about records and allowing confidence thresholds to be tailored to fit specific research aims' (Ranade 2016).

⁴⁰⁰ Video presentation by Sonia Ranade 2016, <https://media.nationalarchives.gov.uk/index.php/traces-time-new-tool-finding-linked-records-across-collections> (Accessed 12 October 2024).

⁴⁰¹ 'As a digital archive creating a digital infrastructure for research, we do not believe that it should be our role to pre-empt the researcher's level of interest in the data we hold. Our preferred model would be to give the researcher control to tailor thresholds to their own purpose – many will want only high confidence matches while others may wish to explore more possibilities' (Ranade 2016).

suggestion at the individual record level,⁴⁰² and most researchers focus attention on individuals or small sets of individuals.

Reflecting on the outcome of the project, the team leader, Sonia Ranade, recognised that digitisation produces a demand for a different, and more detailed, kind of affordance for researchers seeking information contained in Records which is different to that garnered by archivists managing collections through metadata.⁴⁰³ It became apparent that the new TTT project probabilistic and multi-source validation approach, coupled with exhaustive provenance data trail requirements, was unsuited to Discovery with its focus on retrieving metadata on one set of linked archival documents at a time.⁴⁰⁴ The TTT team reflected that archival records, because they have evolved over time and were not originally designed to be a part of a digital affordance, are poorly suited to provenance tracking to sources: ‘We are moving from a world in which archival data is metadata and the value of the record is locked away in paper scanned images or to one where the whole record can be mined and analysed’ (Ranade 2016).⁴⁰⁵ Concerns arose at the micro level too. Even attempting to make linkages between data in a variety of sources using person name as the link access points

⁴⁰² 27 October 2023. Meeting with Mark Bell, senior digital researcher, TNA.

⁴⁰³ ‘Our current descriptive standards do not offer a sufficiently pragmatic or flexible approach to describing the range of material we are receiving and the task of applying current descriptive practice to the “digital heap” is rapidly becoming a barrier to the transfer and accession of digital records’ (Ranade 2016).

⁴⁰⁴ ‘Firstly, these links are neither curated nor authoritative. They are automatically generated at scale and are not individually checked. Secondly, they are not bald assertions: every link is qualified with an indication of our confidence in that link. And because we can apply different linking methods at different times to generate multiple links with multiple confidence measures, we must also hold information about the provenance of each link. This work has created data of a size and shape that will not “fit” into a first-generation digital archival catalogue such as Discovery’ (Ranade 2016).

⁴⁰⁵ ‘[W]e are seeing a shift in the nature of the archival file and a need to manage and provide access at a lower level of granularity: to individual digital objects, instead of papers collected together under a physical cover. It is becoming unsustainable for our archivists to facilitate access to these records through creating individual authoritative, high-quality archival descriptions. We are moving from a world in which archival data is metadata and the value of the record is locked away in paper scanned images or to one where the whole record can be mined and analysed. Whilst curation practices place great emphasis on the provenance of our records, we do not generally capture provenance for our metadata. This is a fundamental shift from our first-generation lists and images, to a second-generation probabilistic and temporally aware platform for managing and publishing archival data’ (Ranade 2016, n.p.).

proved challenging, because existing archival ontologies were perceived to be inadequate to the task (see Chapter 4).⁴⁰⁶

5.5 ResearchSpace

5.5.1 Background

ResearchSpace began as a project embedded in British Museum curatorship in 2010 with a small seed grant from the Mellon Foundation.⁴⁰⁷ In 2014, after a three-year development period and further substantial funding from the Mellon Foundation, the project moved from a conceptual to a development phase.⁴⁰⁸ ResearchSpace was used by the British Museum to perform several GLAMS expert research projects, for example ‘Late Hokusai: Thought, Technique, Society’⁴⁰⁹ and ‘Ancestors, artefacts, empire – mobilising Aboriginal objects’.⁴¹⁰ Each project had its own team members, with Dominic Oldman, head of ResearchSpace, involved as Co-Investigator. In November 2021 ResearchSpace was adopted by TNA as an in-house research tool.⁴¹¹

⁴⁰⁶ ‘The concept of the “person” is at the heart of this project, and we required a data architecture that could potentially encode information about every individual who appears in a public record. There are published schemas for person however, these will neither accommodate the variety of identifying data attributes encountered in our records, nor encode the fuzziness that is a feature of archival data’ (Ranade 2016).

⁴⁰⁷ 100,000 USD June 2010 (8 month project): <https://www.mellon.org/grant-details/researchspace-8582> (Accessed 7 September 2024).

⁴⁰⁸ 1,500,000 USD March 2014 (2.5 year project): <https://www.mellon.org/grant-details/researchspace-10431> (Accessed 7 September 2024).

⁴⁰⁹ From 2016 to 2019 (10–15 team members): <https://www.britishmuseum.org/research/projects/late-hokusai-thought-technique-society> (Accessed 7 September 2024).

⁴¹⁰ Four phases of work 2011–2015, 2016–2020, 2016–2021 and 2016–2023, 20–25 team members: <https://www.britishmuseum.org/research/projects/ancestors-artefacts-empire-mobilising-aboriginal-objects> (Accessed 7 September 2024).

⁴¹¹ ‘The National Archives (TNA) has implemented a new institutional digital system in its Collection Care department using ResearchSpace, a system that captures the knowledge and processes of cultural heritage

In 2023 ResearchSpace as an entity moved from the British Library to Kartography.org⁴¹²

with Dominic Oldman as director.⁴¹³ As at the time of writing, the adopters of

ResearchSpace, in addition to the British Museum and TNA, include:

- I Tatti – The Harvard University Center for Italian Renaissance studies.⁴¹⁴
- The Linked Infrastructure for Networked Cultural Scholarship (LINCS).⁴¹⁵
- The Pharos ResearchSpace platform.⁴¹⁶
- Members of the CORDH⁴¹⁷ community use ResearchSpace technology for projects and ongoing research. They include the two Max Planck Institutes (History of Science and Bibliotheca Heriziana), the University of Zurich and ETH University.
- Linked Conservation Data (Velios Athanasios and St. John Kristen 2021). A network of twenty-three partners, led by the University of the Arts London and Stanford University, including the Bodleian Library, Library of Congress, Fitzwilliam Museum, Getty Research Institute, National Gallery London and The National Archives.

ResearchSpace has a commercial/research development relationship with ‘metaphacts’,⁴¹⁸ a leading commercial digital solutions provider, with research offerings across pharma and life

and humanities professionals and using the CIDOC CRM (Conceptual Reference Model). This new system records both conservation practice and research with provenance and historical context. It provides experts with a flexible and expandable information system that dynamically grows as processes change.’
https://researchspace.org/blog/national_archives_researchspace (Accessed 7 September 2024).

⁴¹² <https://kartography.org/about.html> (Accessed 7 September 2024).

⁴¹³ <https://researchspace.org> (Accessed 7 September 2024).

⁴¹⁴ <https://www.hup.harvard.edu/series/the-i-tatti-renaissance-library> (Accessed 7 September 2024).

⁴¹⁵ <https://lincsproject.ca/docs/about-lincs> (Accessed 7 September 2024).

⁴¹⁶ <http://pharosartresearch.org> (Accessed 7 September 2024).

⁴¹⁷ <https://www.cordh.net> (Accessed 7 September 2024).

⁴¹⁸ ‘Knowledge graph technologies have become prominent in the context of the cultural heritage domain, where the CIDOC-CRM Ontology became a popular standard for exposing cultural heritage information as linked data. The metaphactory platform is utilized in the context of the ResearchSpace project to manage the British Museum knowledge graph and help the researchers (a) explore meta-data about museum artifacts: historical context, associations with geographical locations, creators, discoverers and past owners, etc, and (b) use this meta-data in collaborative work by creating annotations, narratives involving semantic references, and argumentations exploiting knowledge graph data as evidence’ (Haase et al. 2019, 12).

sciences, engineering and manufacturing, and cultural heritage.⁴¹⁹ It was not possible to discuss this relationship with the commercial partner Kartography. However, it is clear from a desktop analysis that metaphacts provides the technology for ResearchSpace through its metaphactory Knowledge Graph⁴²⁰ and its other considerable cultural heritage offerings see Figure 5.15).⁴²¹

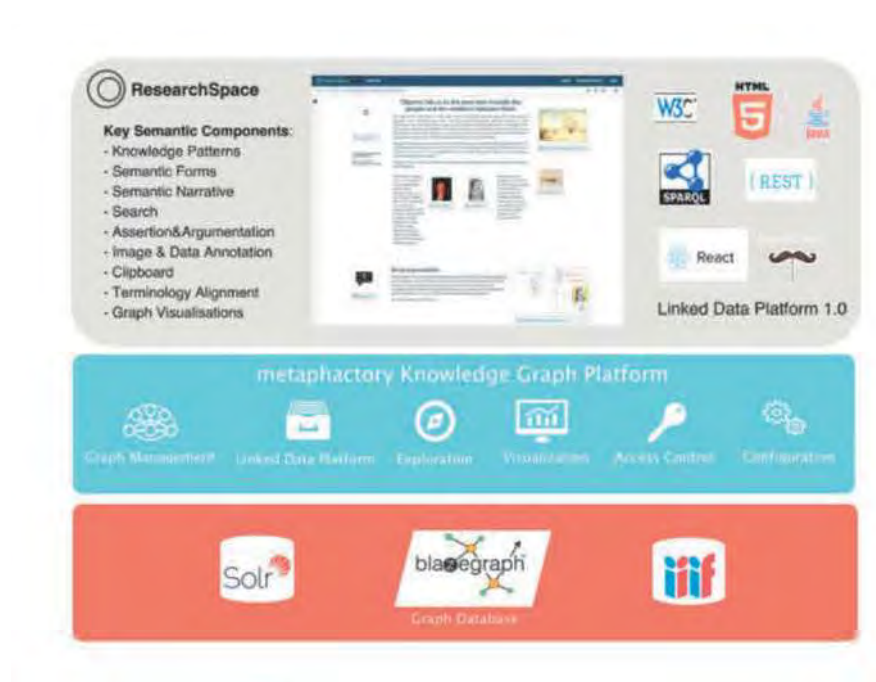


Figure 5.15 ResearchSpace platform architecture showing metaphactory as the comprehensive platform (Oldman and Tanase 2018, 333)⁴²²

⁴¹⁹ 'metaphacts is a Germany-based company delivering metaphactory – a platform that empowers customers to accelerate their knowledge graph journey and drive knowledge democratization, improve data literacy and reach smarter business decisions with data.' <https://metaphacts.com/company/about-us> (Accessed 7 September 2024).

⁴²⁰ 'metaphactory, a platform for building knowledge graph management applications. The metaphactory platform aims at supporting different categories of knowledge graph users within the organization by realizing relevant services for knowledge graph data management tasks, providing a rich and customizable user interface, and enabling rapid building of use case-specific applications' (Haase et al. 2019). An introduction to the ResearchSpace application of the metaphactory knowledge graph can be viewed at <https://www.youtube.com/watch?v=MaAv0SE7wis&t=13s> (Accessed 7 September 2024).

⁴²¹ <https://metaphacts.com/solutions/cultural-heritage?semanticSearch-search1=N4lgziBcoNZQbAGhANYiEyAmB7dYBjACwFMBbAQ0gGEAnEigFwEsUSB1HWuAXx6A> (Accessed 7 September 2024).

⁴²² See also (Oldman and Tanase 2018, 332): 'The ResearchSpace technology stack builds on the metaphactory knowledge graph platform enabling customisation and extensibility of the interaction with the graph database (Blazegraph7) through the use of familiar open standards such as RDF and SPARQL, expressive ontologies for

5.5.2 The ResearchSpace canvas

ResearchSpace supports three different types of search scenarios. Each differs in the way in which the system handles the formulation and transformation of a query into a set of resources. These are:

- (a) knowledge-graph driven search,
- (b) knowledge pattern-based search, and
- (c) text-based search. (Oldman and Tanase 2018, 335)

Dominic Oldman and Diana Tanase define ResearchSpace as an ambitious Semantic Web offering for interdisciplinary digital research.⁴²³ Oldman and Tanase characterise the ResearchSpace affordance as a ‘thick’ system because it dynamically combines epistemology with ontology to provide a wide scoped service.⁴²⁴ ResearchSpace is a more complex system (in terms of both data and the assembled technologies) than SNAC, and is considered by Oldman and Tanase as a significant improvement on most other ‘thin’ affordances in this

schema modeling based on CIDOC CRM, rules, constraints, and query specifications based on SPIN, W3C Web Components, W3C Open Annotation Data Model, and W3C Linked Data Platform Containers. The platform is open source, integrating external tools including OntoDia, MIRADOR Image Viewer with an IIIF Image Server. Instantiating ResearchSpace for application projects involves creating templates, which are a mixture of HTML5, React Components and Handlebars.’

⁴²³ ‘A Semantic Web knowledge oriented system that is designed to work in, or help transform, knowledge environments into collaborative, argumentative, digital scholarly spaces through the contextualisation of data using ontoepistemological processes for semantic modeling. It supports interdisciplinary research and is additionally underpinned by material culture representing world history through the products of social relations’ (Oldman and Tanase 2018, 339).

⁴²⁴ ‘A “thick” information system [is one] using structured data in which there are flexible and expandable structures of information supporting different interdisciplinary vantage points describing internally related processes [entities] with explicit semantics and context, and where processes can be connected across different types of time and space, whether absolute, relative or relational’ (Oldman, Tanase, and Santschi 2019, No page numbers).

field.⁴²⁵ ResearchSpace is in essence a knowledge graph desktop application called a canvas with an impressive array of functions and applications⁴²⁶ (Figure 5.16).

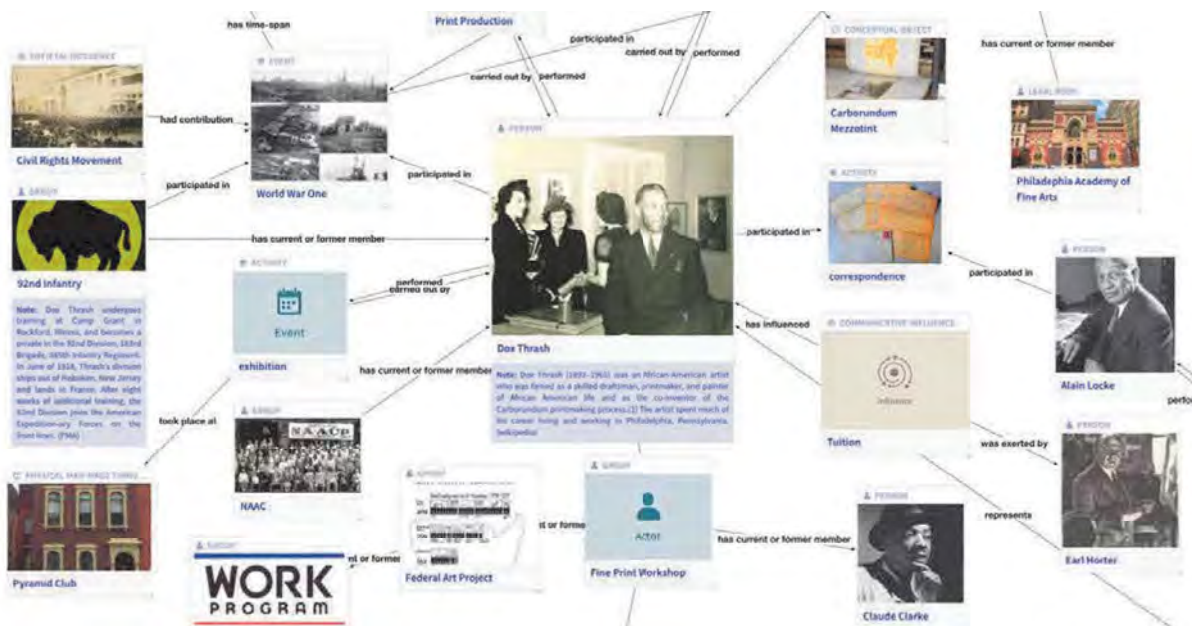


Figure 5.16 A ResearchSpace knowledge graph represents data in a network of meaningful relations (<https://researchspace.org>, accessed 7 September 2024)

ResearchSpace’s metaphactory concept of semantic modelling has three overlapping layers which comprise the ResearchSpace framework typology:

- The mapping of instances of data into the metaphactory platform using automated reasoning, allowing automated and human interventions for interpretation and execution.

⁴²⁵ ‘A “thin” information system [is one] that stores and processes structured data with a predefined data model, used to record independent instances of entities with little or no explicit semantics or contextualised relationships, typically presenting information in absolute time and space for the purpose of creating a finding aid or essential reference’ (Oldman, Tanase, and Santschi 2019, No page numbers).

⁴²⁶ <https://researchspace.org/semantic-tools> (Accessed 7 September 2024).

- The consistent representation of terms, concepts and relationships, with shared domain-specific understanding to reduce ambiguity so that public vocabularies and ontologies can be imported and used.
- An ontological layer to allow concepts and relationships to be defined and determined at the data instance level by using standardised machine readable ontologies and vocabularies, resulting in human and machine readable and operable data instances facilitating actionable insights and contextual analytics (Figure 5.17).

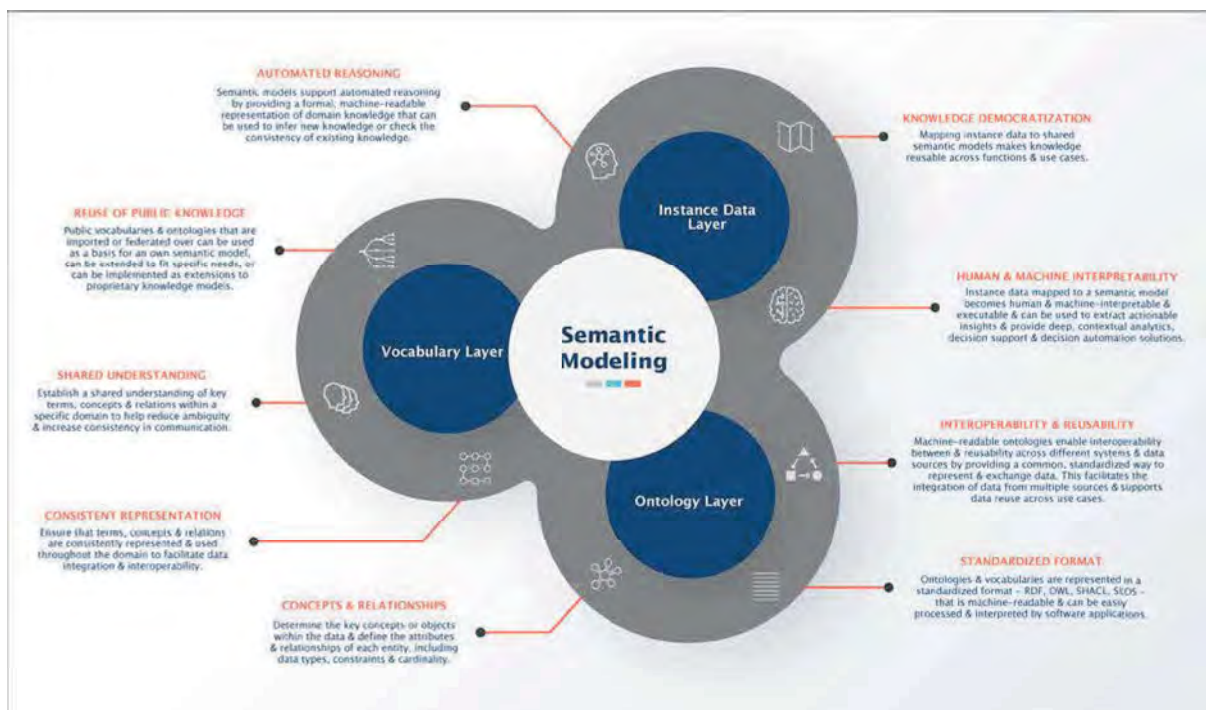


Figure 5.17 The three layers of a semantic knowledge graph (<https://blog.metaphacts.com/importance-of-semantic-knowledge-graph>, accessed 7 September 2024)⁴²⁷

⁴²⁷ 'A semantic knowledge graph is a large network of entities representing real-world objects, like people, organizations and abstract concepts, such as professions and their semantic relations and attributes, through a visual graph structure. While many other varying definitions exist, our definition of the knowledge graph places emphasis on defining the semantic relations between these entities, which is central to providing humans and machines with context and means for automated reasoning.'

<https://blog.metaphacts.com/importance-of-semantic-knowledge-graph> (Accessed 7 September 2024).

The ResearchSpace canvas can produce a rich user interface experience using imported datasets of the user's choice (providing they conform to the metaphactory data, vocabulary and ontological model). (See Figure. 5.18.)

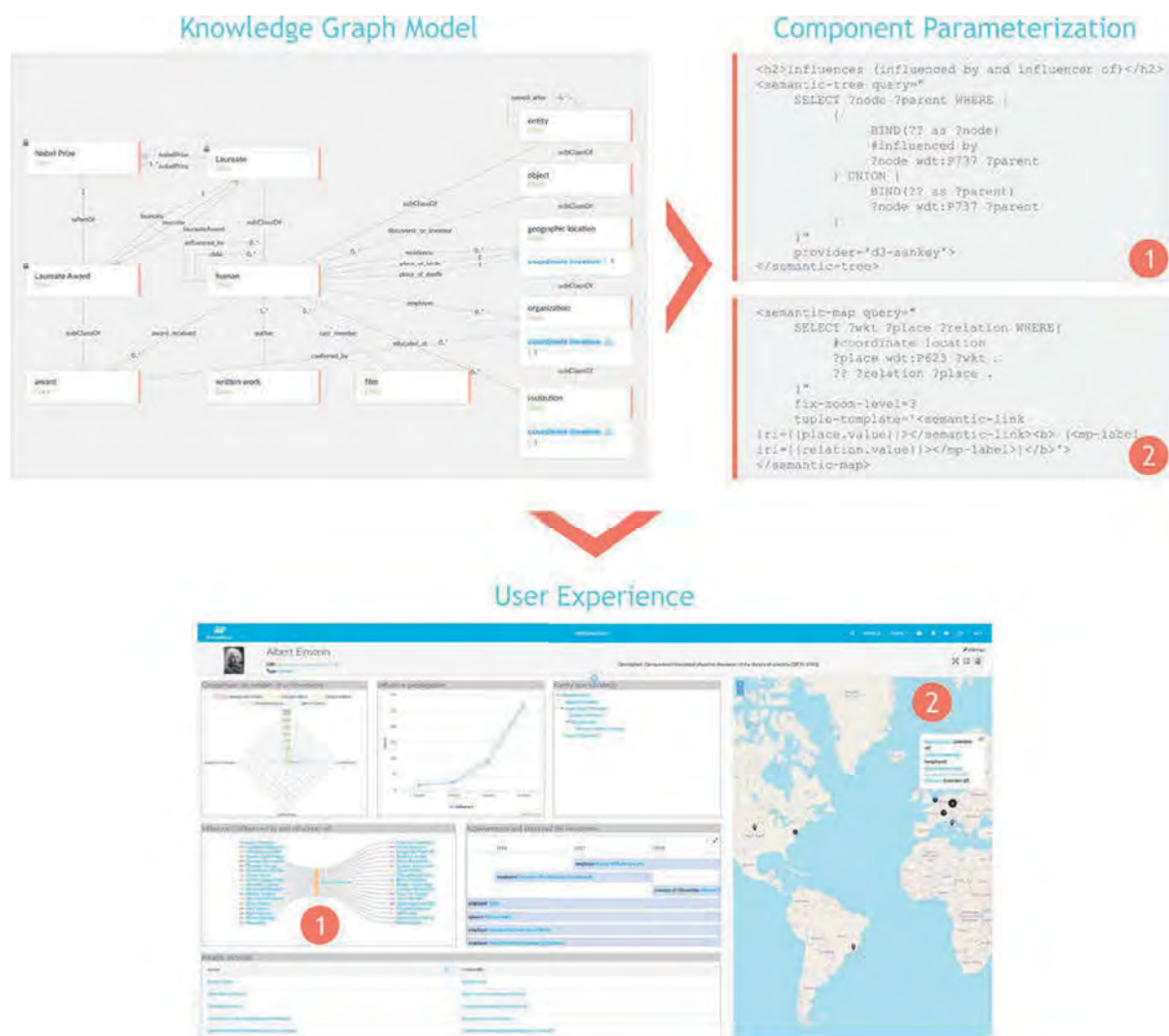


Figure 5.18 A user interface in metaphactory (<https://blog.metaphacts.com/visual-ontology-modeling-for-domain-experts-and-business-users-with-metaphactory>, accessed 7 September 2024).⁴²⁸

⁴²⁸ 'We included detailed configuration examples for two components in the user interface: A Sankey diagram which highlights how one person – in this case, Albert Einstein – influenced and was influenced by other persons. An interactive map which displays locations related to important events in this person's life. As the

5.5.3 EBP in ResearchSpace

ResearchSpace from its original conceptual design recognised the importance of EBP, especially when data is used for relationship building and analysis. The development team early on acknowledged this as a primary concern based on the team's understanding that earlier projects of this kind had largely failed to fully grasp the need for robust semantic interoperability, essential for relationship building.⁴²⁹ The ResearchSpace team also recognised the importance of evidenced-based data, fixity, affixedness and provenance in the digital representations of physical items.⁴³⁰ The ResearchSpace vision and objectives align well with those of this thesis, the need for EBP as the next phase of the development of digitisation in DH.⁴³¹ Dominic Oldman, Martin Doerr and Stefan Gradmann (the ResearchSpace team) identified four major issues to which EBP is an appropriate response:

In the Linked Data world, we therefore have four major issues:

- We need to differentiate between 'known' facts and 'possible' facts.

ontology evolves and new datasets are connected, the user experience will automatically update to reflect these changes in the model and in the data.' <https://blog.metaphacts.com/visual-ontology-modeling-for-domain-experts-and-business-users-with-metaphactory> (Accessed 7 September 2024).

⁴²⁹ 'Despite strong examples of the use of relational databases in the digital humanities, particularly in the area of prosopography, lack of syntactic and semantic interoperability has inevitably limited the ability of structured data projects to reach beyond relatively narrow scopes, and has arguably contributed to a fragmentation of information and an accumulation of siloed [even if "linked"] data repositories' (Oldman, Doerr, and Gradmann 2015, 254).

⁴³⁰ 'A digital representation must first and foremost provide a faithful, understandable, and explainable representation of a source as a basis for further valid scholarly investigation' (Oldman, Doerr, and Gradmann 2015, 256).

⁴³¹ 'This was impossible in the past, and is a new "innovative" ability digital humanities can provide. By representing the implicit relationships embedded in institutional datasets, an opportunity exists to establish a knowledge base that is both rich and broad enough to fuel more sophisticated digital humanities methods supported by numerous and varied historical perspectives. Collaboration with memory institutions on this single issue of digital data curation could dramatically improve the quality of humanities research, with wide-ranging benefits for society' (Oldman, Doerr, and Gradmann 2015, 260).

- We require a model of nested [as opposed to flat] relationships, to provide the possibility of integrating data that properly represents the scholar's knowledge.
- We need to provide information with a description of reality to the level that allows us to participate in meaningful discourse at any level.
- We must always be able to trace the provenance of knowledge back to the source micro-level [with its original context and perspective intact]. (Oldman, Doerr, and Gradmann 2015, 260)

The first concern calls for discernment in information contained in sources. The EBP system requires that this be evidential and prosopographical. The ResearchSpace team refer to 'facts' (see the discussion in Section 1.4 for the difference between facts and evidence).

The second concern, that researchers naturally work with nested relationships, is addressed by the NAI-UID system, which links data through hierarchical relationships, because these occur naturally in the prosopographical data on PHL. Other attributes can be added, such as occupations and locations, to further develop the web of relationships naturally found in the Records.

The third concern, that data needs to represent 'reality', is at the core of the EBP system with its focus on the information contained in Records.

The fourth concern, for tracking data back to its Record, is also at the heart of the EBP system, and is addressed in the systematic application of UIs to incidences of EBP – the Record containing the incidence, the collection the source can be found in, the archive where the collection resides, and the authorised name standard the incidence is attributed to.

Although ResearchSpace has been adopted by several cultural heritage institutions, there is no evidence of its use beyond institutions. Perhaps it is the case that affordances like

ResearchSpace can/will only be used within large institutions that can support its use.

ResearchSpace needs direct access to several very large data records for its use to be efficient. Oldman and Tanase recognise that unless institutions make their large datasets fully accessible to the public (requiring non-academics to be trained in the structures, features and particularities of institutional datasets, and the complex ResearchSpace suite of technologies needed to exploit them), ResearchSpace will have little practical use beyond the firewalls of institutional data systems.⁴³² A possible future in which a kind of elitism emerges with only curators, archivists and librarians able to fully benefit from large data affordances and their technologies should be avoided.⁴³³ Another possible future sees general researchers wishing to work with large datasets restricted to a data service that gives some limited access to institutional large datasets, with data accessed through user-friendly simplified technological arrays.⁴³⁴ Unfortunately this will not provide systematic access to Records, though it will point to the location of those Records. Neither of these possible futures will maximise the benefit of digitisation in DH, or be a suitable alternative to the EBP system.

⁴³² 'In the last decade several museums including the British Museum have opened their collection data to the World Wide Web. Yet, this is problematic since cultural heritage collection data systems were designed for internal administration by specialist users, where the shortfall in data specificity, ambiguities, or uncertainties are compensated by the knowledge of expert users who interpret it. The language and the knowledge required to understand the original meaning behind the data is not accessible to external users when this data is openly published in Linked Data format' (Oldman and Tanase 2018, 327).

⁴³³ 'Traditionally, digital humanities projects have mostly crafted their own datasets limited by the resources available to any individual project. While the research questions they addressed have been useful and informative, projects lack the ability to call upon larger repositories, despite the significant amounts of accumulated data created by the large investments in digitalization on the part of memory institutions over the last 30 years. This has led to criticisms that research projects concentrate disproportionately on the technology rather than on the content they analyse and the scope of questions they address' (Oldman, Doerr, and Gradmann 2015, 264).

⁴³⁴ 'The investment of large amounts of money in one-size-fits-all harvesting mechanisms, and then converting this to Linked Data remove much of its original value and provides no correspondence to original knowledge. This seems to go against the very spirit and nature of why Linked Data and Semantic technologies were created, in which enfranchisement is a key goal' (Oldman, Doerr, and Gradmann 2015, 265).

5.5.4 A critique of ResearchSpace – the Linked Conservation Data project

The Linked Conservation Data (LCD) project,⁴³⁵ which began in 2019 and ran for two years, was undertaken by a consortium of twenty-three major archives.⁴³⁶ The project's objective was to assess issues in the use of digital conservation documentation records for research and the project focused on three key areas – terminology, modelling and dissemination.⁴³⁷

The project team used ResearchSpace as its core research tool for undertaking the pilot study.⁴³⁸ The study was extensive, comprising eleven international workshops and assessing hundreds of conservation-controlled vocabularies.⁴³⁹ The project produced two reports, a Terminology Report in 2019 and a Board Reattachment Pilot Report in 2022.

⁴³⁵ 'Linked Conservation Data is a Network of partners working on improving access to conservation documentation records. The aim of the Network is to discuss and report on ways that conservation documentation can be disseminated and re-used more effectively through Linked Data.'
<https://www.ligatus.org.uk/lcd> (Accessed 12 October 2024).

⁴³⁶ <https://www.ligatus.org.uk/lcd/consortium> (Accessed 12 October 2024).

⁴³⁷ 'We have identified three areas of development for the network's attention: Terminology, Modelling, and Dissemination. Terminology: In the Semantic Web, communicating by using a variety of terminology traditions is important for disambiguation. The Network will assess the suitability of existing vocabularies in conservation and identify the amount of work needed both in terms of coverage and in terms of formatting to improve them for use in Linked Data applications. The relevant Linked Data standard for vocabularies is SKOS. Modelling: In the Semantic Web, the type of each published record needs to be explicitly declared. For example, machines need to be able to handle records of type condition assessment and records of type treatment proposal differently. A standard which provides different types of records (classes) is the CIDOC-CRM. The Network will assess the suitability of the CRM and its extensions for conservation. Dissemination: The Network will share best practices for producing Linked Data from conservation documentation and report on the readiness and capacity of existing software to host and share Linked Data.'
<https://www.ligatus.org.uk/lcd> (Accessed 12 October 2024).

⁴³⁸ 'The steps undertaken to implement the pilot on ResearchSpace are presented here sequentially, but in some cases iterations of these steps were required:

Imported transformed data.

- Modified ResearchSpace templates to add extra details.
- Imported photographs and linked them to items.
- Built search pages for the research questions.
- Built narratives based on queries in the data.

The above consisted of technical work which depends on familiarity with the ResearchSpace architecture and a querying language for RDF called SPARQL' (Velios Athanasios and St. John Kristen 2021).

⁴³⁹ <https://www.ligatus.org.uk/lcd/controlled-vocabularies> (Accessed 12 October 2024).

The Terminology Report set out that the team adopted the Arts and Architecture Thesaurus (AAT) Getty Vocabulary as its core thesaurus, seeking to conform other vocabularies with this to facilitate interoperability of data.⁴⁴⁰ The complexity of this task was explained both in terms of the variability of data⁴⁴¹ and diverse terminologies,⁴⁴² necessitating the production of alignment data support advice, which the team agreed to publish online to enable users of the data to understand the complex data management actions undertaken.⁴⁴³ Worryingly (for a large project of this size and with considerable support from many prestigious institutions), the team was only able to promise that ‘It will be hosted by a consortium member for as long as funding is available’ (Velios 2019, 3). This is another example of the precariousness of the long-term viability of DH projects.

The second report describes a joint exercise between the project members to integrate information on a shared activity,⁴⁴⁴ an LOD project on board attachment preservation

⁴⁴⁰ ‘The AAT (<http://www.getty.edu/research/tools/vocabularies/aat/>), part of the Getty Vocabularies programme, offers extensive, but not complete, coverage of conservation terminology. We decided that the AAT will be a core thesaurus for the project’ (Velios 2019, 2).

⁴⁴¹ ‘The current state of conservation glossaries/vocabularies is diverse in terms of readiness to be used in Linked Data using SKOS. Few vocabularies are published as Linked Data or offer unique identifiers for the concepts/terms that they include. Some are published as structured or semi-structured data which will require relatively limited work to turn them into Linked Data. Some are only available in unstructured text format or published in print and would require significant effort to share as Linked Data’ (Velios 2019, 2).

⁴⁴² ‘Conservation records are produced in different institutions using different terminology. Searching across different records often requires an understanding that a term in one record and a different term in another point to the same concept. Therefore we consider aligning concepts/terms and vocabularies (reconciliation) an essential task for this project. SKOS provides relevant tools for expressing how concepts are aligned. We have decided that we will accommodate alignment data between any two vocabularies. However given the unique position of the AAT in the eld we encourage alignment with the AAT in addition to any other direct alignment’ (Velios 2019, 2).

⁴⁴³ ‘LCD will produce and maintain vocabulary alignment data. This includes any statements using the SKOS equivalence (close or exact match), associative (related concept) and hierarchical (broader concept) relationships across different vocabularies. This dataset will be held centrally in a repository with a distributed management team which will include members from the LCD consortium’ (Velios 2019, 3).

⁴⁴⁴ ‘The pilot created Linked Data for a common conservation treatment method from the records of multiple institutions in order to develop workflows, identify challenges and generate recommendations for future Linked Data use in Conservation. The pilot project steps were: developing research questions, selecting records, aligning terminology, modeling the data, and uploading to a portal. This report contains detailed analysis of each step including results, performance and recommendations for future pilots or Linked Data Implementations’ (Velios Athanasios and St. John Kristen 2021).

activities in archives (necessary when board covers to old or damaged books are themselves damaged or become detached).⁴⁴⁵

The primary data from the data donors consisted of structured and semi-structured data in spreadsheet format, individual DOCX files and paper Records,⁴⁴⁶ some two hundred records in total.⁴⁴⁷ The records were converted into machine readable format, and this was the most time-consuming part of the whole project. It was later felt that variability in data provision and formatting was a significant barrier to the possible scaling up of the project to accept data at large volumes.⁴⁴⁸ Using the outcome of the Terminology Project Vocabulary Analysis standardised data was then extracted and further difficulties were discovered – that some terms were not present in the Getty AAT, and that some of the terms used by archivists many years ago were now obscure.⁴⁴⁹ This is clear evidence of this thesis's concern when

⁴⁴⁵ '[A] pilot implementation of Linked Data in conservation using a specific case study from the field of book conservation: board re-attachment. Books with detached boards are common in historic library collections and conservators have applied a variety of techniques and materials when re-attaching boards over the years. Data from four consortium members were included in the pilot. These were: the Bodleian Library, the Library of Congress, The National Archives (UK) and Stanford Libraries. Partner members from the British Museum (Dominic Oldman, Cristina Giancristofaro, Diana Tanase) contributed to the meetings related to the pilot portal.' <https://lcd.researchspace.org/resource/rsp:Start> (Accessed 15 July 2024).

⁴⁴⁶ 'Library of Congress: structured data, available in spreadsheet format.

Bodleian Library: free text conservation reports, text-based documentation forms on paper.

The National Archives: semi-structured data, available in spreadsheet format.

Stanford Libraries: scanned reports, structured data forms in individual .docx files' (Velios Athanasios and St. John Kristen 2021).

⁴⁴⁷ 'Each partner contributed about 30–50 records covering the time-span of 50 years with the Library of Congress contributing a large number of records from recent years, reflecting recent staff efforts to organize information on treatments in a spreadsheet' (Velios Athanasios and St. John Kristen 2021).

⁴⁴⁸ 'Manually converting text-based records from individual conservation reports to structured records was the most time-consuming aspect of this task. This is not scalable for legacy data and the recommendation is that, at least, new records are produced in a digital format following a schema within a database system. For legacy data, such conversion would need to happen at metadata level with enough detail to point to the full reports' (Velios Athanasios and St. John Kristen 2021).

⁴⁴⁹ 'Many terms were not included in the Getty AAT and therefore could not be aligned with it. This meant that two different terms in use in two different partners but pointing to the same concept could not be jointly queried if they were not included in AAT. Our plans for submitting these terms to the Getty AAT were not materialised due to the overall delay of this task and the requirement for extensive metadata accompanying newly submitted terms to the Getty AAT. A local hub thesaurus could have been created to accommodate such terms but we decided against it so that we can draw attention to the importance of alignment of local terms. Terms used in local vocabularies in previous decades become obscure when members of staff retire and

building digital datasets from archival metadata about too easily trusting data extracted from old archival Records made in the pre-digital era. The project team found that using ResearchSpace to perform the task helped them to intervene locally and make modifications to enable data to be assimilated.⁴⁵⁰ The principal learning from the pilot study was that:

- Partners found the project challenging, they lacked the time commitment and necessary resources, and they had poor understanding of the work of the team.⁴⁵¹
- The CIDOC-CRM was very difficult to manage because of its complexity.⁴⁵²
- The team found ResearchSpace an excellent tool for the project but noted that adapting the tool for the project was a job for a ‘technical expert’ and that due to the technical complexity of ResearchSpace, partner engagement was poor.⁴⁵³

the pattern of use of these terms is no longer recognised. Partners highlighted the value of having such terms documented through local scope notes and alignment with externally maintained reference thesauri’ (Velios Athanasios and St. John Kristen 2021).

⁴⁵⁰ ‘The implementation of the pilot on the ResearchSpace platform meant that a set of tools provided out-of-the-box could be customised and used for cross-searching terminology within the limited scope of the pilot. The flexibility offered by ResearchSpace allows planning of complex tools for querying vocabularies and allowing versioned control of their alignment with hub thesauri. A domain expert vocabulary hub for conservation is a significant task that will benefit future projects’ (Velios Athanasios and St. John Kristen 2021).

⁴⁵¹ ‘The plan of the project made provisions for partner engagement at all stages of the transformation of the datasets to Linked Data datasets. Such engagement was limited during mapping the datasets to the CIDOC-CRM. The main factor for this was that some partners found the ongoing terminology tasks challenging or time-consuming and did not have the resources available to take on additional complexity around modelling. Some partner feedback points to the lack of engagement during this part of the project. Also, keeping the modelling team small among project members with significant experience with modelling meant that we could undertake the work faster but without communicating our experience to partners as inclusively’ (Velios Athanasios and St. John Kristen 2021).

⁴⁵² ‘The CIDOC-CRM offers classes and properties, the number of which often overwhelms newcomers. CIDOC-CRM classes and properties alongside those from its extensions cover a seemingly limitless number of scenarios, and this proves challenging to those getting started in modelling as well those introducing them through didactic material. The profile should be accompanied by concise, non-technical, and domain-specific training material to help lower barriers to entry for those wishing to adopt systems that use the CRM’ (Velios Athanasios and St. John Kristen 2021).

⁴⁵³ ‘Importing records and building queries can be considered as the ultimate test for integration as querying and returning results can prove the success of the project. It was possible to encode the pilot research questions using SPARQL queries and present the results in the form of timelines and diagrams which assisted with the comprehension of the answers. Existing familiarity with the ResearchSpace platform and support from the British Museum as part of their role in the project, meant that this work was put in place efficiently.

- The team did not make as much use of ResearchSpace as they desired due to its complexity. They thought ResearchSpace not yet fully developed.⁴⁵⁴

Accepting the limitations noted above, the team had a favourable opinion of ResearchSpace as a tool for GLAMS personnel to use in managing and analysing archival materials.

However, the review of the LOD project suggests that ResearchSpace is too complex to find wide appeal outside of the small group of technical experts found in GLAMS. The

ResearchSpace team noted (the four concerns, see Section 5.5.3) the affordance limitations with regard to EBP and therefore its use for research into PHL.

Narratives from conservation reports enriched with results from the encoded data were also produced using ResearchSpace. Apart from the British Museum, the University of the Arts London and Stanford Libraries, other partners made limited contributions during this task due to the technical nature of the work. The implementation was demonstrated to the consortium with explanations on its functionality. While the ResearchSpace team aims to make the software easy to use for users without expertise on Linked Data, customising the system and building the required queries remains a job for a technical expert' (Velios Athanasios and St. John Kristen 2021).

⁴⁵⁴ 'ResearchSpace is a significant project in the field of Linked Data which has enabled communities to share data and build new connections and knowledge with them. While the flexibility of the system is evident, allowing great variety of customisations, at the moment the default configuration did not cover all our requirements and we did not have the time and resources to implement them as part of this pilot. These requirements are summarised here:

The default advanced search tool does not automatically scan the models in the available data. This means that a user cannot build their own queries, but has to rely on pre-built queries provided by an administrator. Configuring advanced search to enable custom queries is possible, but required more time than we had available.

The custom querying tool also does not observe the hierarchy logic of the CIDOC-CRM. For example features like class and property hierarchies and property inheritance need to be implemented manually although they are the core of the theory of integration with the CIDOC-CRM.

The default vocabulary manager does not allow querying reconciled concepts across vocabularies. For example, our local vocabularies were reconciled with the Getty AAT, but it was not possible to use that reconciliation automatically when searching terms' (Velios Athanasios and St. John Kristen 2021).

5.6 The Golden Agents

5.6.1 Background

The Golden Agents project⁴⁵⁵ is a ‘large investment program’ that provides a single platform consisting of LOD from several creative archives (see Figure 5.19).⁴⁵⁶ The project’s data comprises records of the producers and consumers of creative goods taken from the digitised Records and the contemporary catalogue records of several collections that contain prosopographical data on persons of the ‘long Golden Age of the Dutch Republic’ (ca. 1580–ca. 1750); these include notarised records of death goods and other related sources of a biographical nature.⁴⁵⁷ The project was developed in the Netherlands from 2017 to 2021 and it was funded by the Dutch Research Council (NWO).⁴⁵⁸ It is managed by a consortium of participating archives⁴⁵⁹ and one of the contributors, LAB1100, provided

⁴⁵⁵ ‘The Golden Agents project (2017–2021), financed by the Large Investments program of the Netherlands Organization of Scientific Research (NWO), is developing a sustainable research infrastructure to study relations and interactions between producers and consumers of creative goods across the long Golden Age of the Dutch Republic (ca. 1580 – ca. 1750). The project will bring together distributed, heterogeneous resources (both existing and new) on creative industries in the Dutch Golden Age as Linked Open Data.’ <https://www.goldenagents.org/about> (Accessed 6 October 2024).

⁴⁵⁶ ‘The consortium of the Golden Agents project consists of institutes of the Royal Netherlands Academy of Arts and Sciences (Huygens Institute for the History of the Netherlands and the Meertens Institute), University of Amsterdam, Utrecht University, VU University of Amsterdam, Rijksmuseum, KB National Library of the Netherlands, City Archives of Amsterdam, RKD Netherlands Institute for Art History and Lab1100.’ <https://www.goldenagents.org/about> (Accessed 6 October 2024).

⁴⁵⁷ ‘The production of art, books, literature and other creative products in this period is covered by many electronic resources like collection databases of museums and libraries, dedicated documentation systems, and research databases. Often biographical records are at the heart of these systems, but typically all these resources use their own subset of biographical data. In the Golden Agents program, we will connect these resources in a linked data framework. This will result in a sustainable infrastructure to study relations and interactions between producers and consumers of creative goods in the Dutch Golden Age’ (Brouwer and Nijboer 2017, 33).

⁴⁵⁸ <https://www.nwo.nl/en> (Accessed 6 October 2024).

⁴⁵⁹ Ja, ik wil!; City Archives of Amsterdam; Notary network; Rijksmuseum Collection; Schrijverskabinet; Amsterdam Corporate Group Portraits; ECARTICO; ONSTAGE; Nederlandse Thesaurus van Auteursnamen; Occasional Poetry; Bredius excerpts; Short-Title Catalogue Netherlands (STCN); Golden Agents HTR – Amsterdam City Archives; Processes of Creativity. <https://data.goldenagents.org/> (Accessed 6 October 2024).

technical expertise and software to the project.⁴⁶⁰ A precursor study to the Golden Agents project was carried out by Brigham Young University (BYU) in 2011.⁴⁶¹ BYU modelled two Historical Social Networks (HSNs) integrated with a genealogical index to the Mormon Migration Database. The modelling was undertaken manually (as here in Project Seven) and linked the two database indexes to the genealogical index (see Section 2.6.2).

⁴⁶⁰ 'LAB1100 is a research and development firm established in 2011 by Pim van Bree and Geert Kessels. LAB1100 brings together skills in new media, history, and software development. Working together with universities, research institutes, and museums, LAB1100 has built the digital research platform nodegoat and produces interactive data visualisations.' <https://lab1100.com/about> (Accessed 6 October 2024).

⁴⁶¹ 'Historical social networks (HSNs) can be used to inform his historical research, including family history and genealogy. In some cases, clues about the structure of an HSN can be found in artifacts of family history such as personal diaries or autobiographical sketches. However, manual inference of such networks can require significant time and effort, including pooling and cross-referencing many different data sources. We present our current research into facilitating that process by automatically finding names in document transcriptions, relating those names to the names found on a roster/list of people who may be talked about in the documents, and automatically generating a social network graph from the result. We link individuals in the social network to a global genealogical database so that people researching their own family histories can easily find their ancestors within the HSNs created in this manner. We also provide examples of how the linked HSNs may be used to inform research about people and situations even when direct information is scarce' (Kennard, Kent, and Barrett 2011, 43).



Figure 5.19 The Golden Agents consortium (<https://www.goldenagents.org/ga-output>, accessed 7 October 2024)

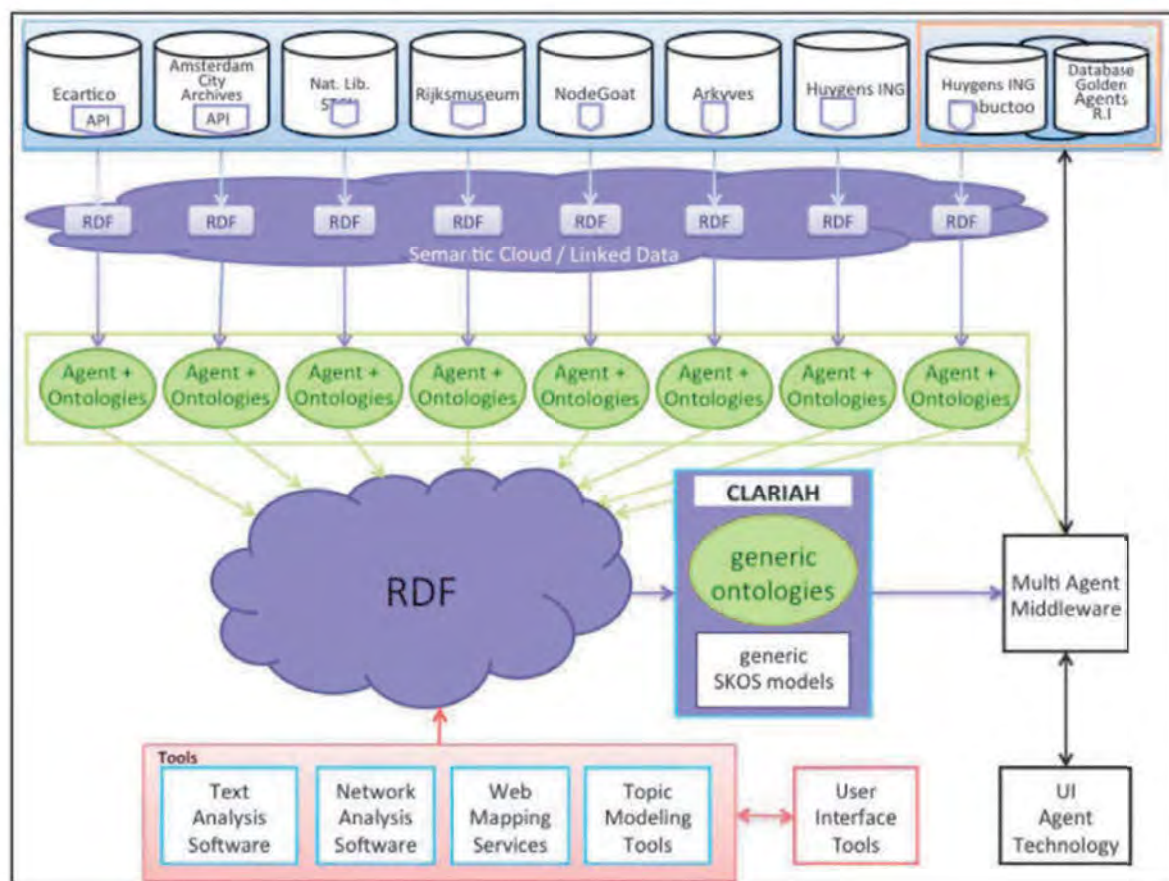


Figure 5.20 The Golden Agents schema – Project Proposal pp 14 <https://www.goldenagents.org/about/>.⁴⁶²

Figure 5.20 shows the eight contributing datasets (top) together with their respective technology providers. Working down the figure, each dataset exists in a variety of formats, and these were first converted to RDF using the relevant ontology for each data source.

Figure 5.21 illustrates the project's attention to granular detail and the focus on extracting all prosopographical data from each source modelled.

⁴⁶² Aanvraagformulier Investerings NWO-groot 2015 BOO Investment Subsidy NWO Large 2015 BOO Application Form. <https://www.goldenagents.org/about> (Accessed 7 October 2024).

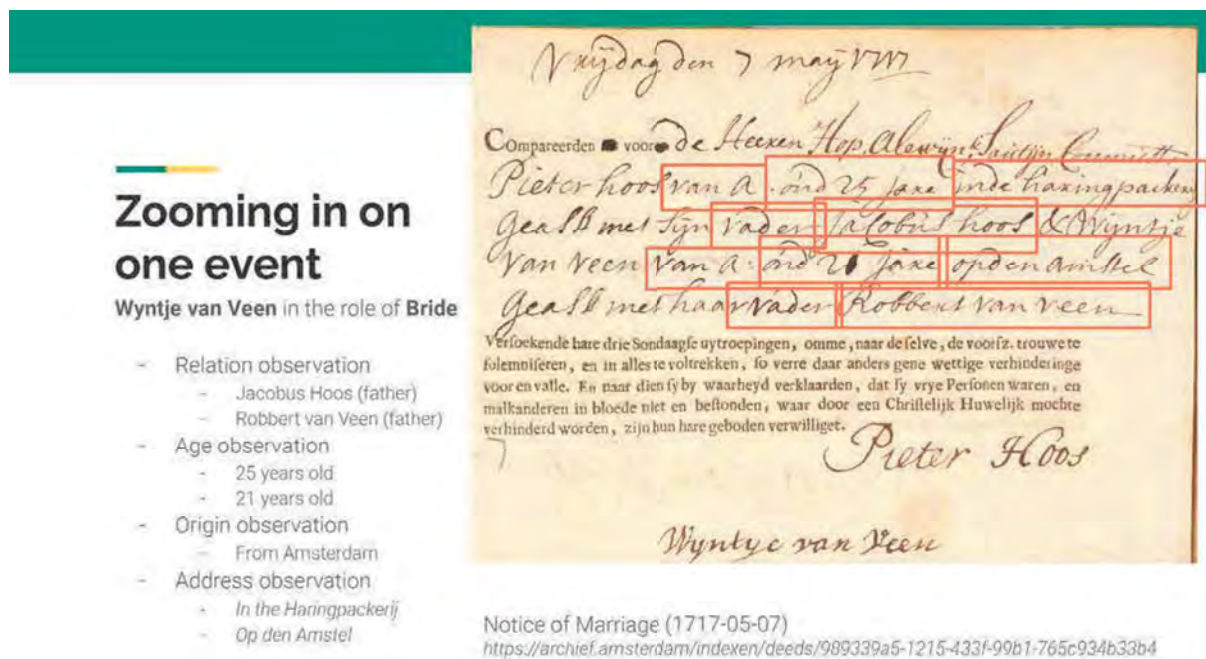


Figure 5.21 The Golden Agents – zooming in on one event
 (<https://www.goldenagents.org/ga-output>, accessed 7 October 2024).⁴⁶³

The converted datasets were then compiled into one combined RDF dataset using the CLARIAH⁴⁶⁴ generic ontologies and SKOS⁴⁶⁵ models, thus bringing all of the data together by using the Reconstructions and Observations in Archival Resources (ROAR) ontology, which the project team modified to create a project specific ontology: ROAR++ (Figure 5.22).

⁴⁶³ Attention to prosopographical data with each observation allocated a URI to enable provenance tracking back to source at all points in the data system.

⁴⁶⁴ <https://www.clariah.nl/organisation> (Accessed 6 October 2024).

⁴⁶⁵ <https://www.w3.org/2004/02/skos> (Accessed 6 October 2024).

Ontology for archival sources: ROAR++

- **Reconstructions and Observations in Archival Resources**
- Model what you **observe**, **interpret** and **infer**
- Then, you make a **reconstruction** based on two or more **observations**

Levels:

1. Collection (cf. **EAD / RICO**)
2. Observation: Content (from **HTR/transcription**)
3. Observation: Direct interpretation (cf. traditional **index**)
4. Observation: Indirect interpretation
5. Reconstruction into one 'individual concept' (cf. **thesaurus/biography**)

Leon van Wissen, Veruska Zamborlini, Challenge LODLAM 2020, Los Angeles, Getty Research Institute, February 3th-4th, 2020 <<https://lodlam.net/challenge-entries/>>

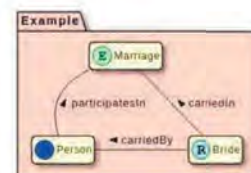


Figure 5.22 The ROAR++ ontology (van Wissen, Zamborlini, and van den Heuvel 2022)

Finally, a set of bespoke tools were designed to interrogate and study the resulting data.⁴⁶⁶

Importantly, the project team recognised early that provenance and disambiguation were essential requirements of the data management and integration process, and that the integrity of the project relied on successfully addressing these two requirements. They effectively solved these problems by using a bespoke application called Lenticular Lens and by adopting the rigorous use of UIs at a granular level (one UID for every individual observation) throughout the project. This facilitated full data provenance by establishing a data trail back to source data, accessible at all points of researcher engagement. The project required considerable data disambiguation and robust data integration.

⁴⁶⁶ 'The Golden Agents research infrastructure enables interaction between various heterogeneous databases by using a combination of semantic web solutions and multi-agent technology that will be supported by ontologies developed together with domain experts. The infrastructure is complementary to, and interoperable with the largest Dutch digital humanities infrastructure CLARIAH. Domain specific ontologies and standards will enrich the generic ontologies and SKOS standards of the CLARIAH infrastructure. Existing data mining techniques, topic modeling methods and network analysis tools of the CLARIAH platform will be offered to researchers to analyze, annotate and visualize metadata and textual/visual sources. Large datasets of images and text can be explored for a much deeper understanding of consumers' responses to developments in styles, genres and fashions in the Dutch Republic.' <https://www.goldenagents.org/about> (Accessed 6 October 2024).

5.6.2 Datasets

Several datasets were collected together in the Golden Agents project (see Figure 5.23).

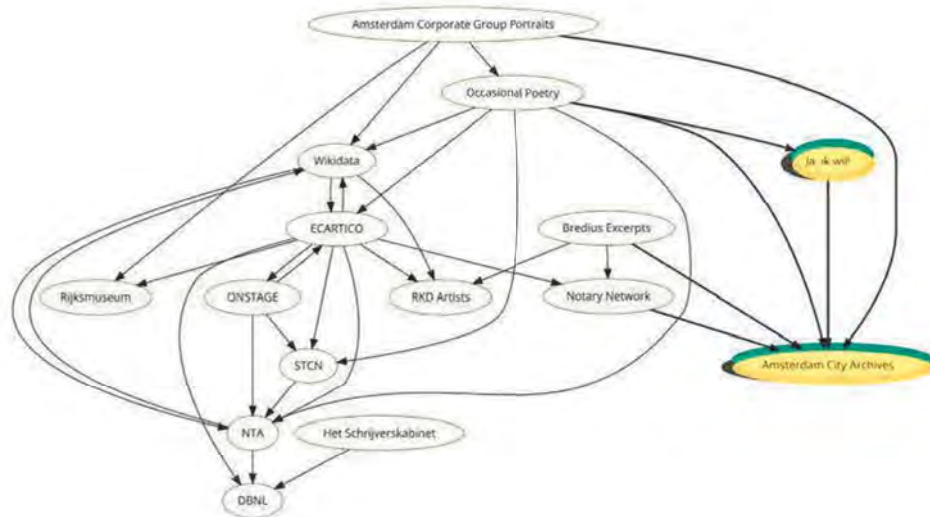


Figure 5.23 The Golden Agents linked datasets universe (<https://www.goldenagents.org/ga-output>, accessed 7 October 2024)

5.6.2.1 ECARTICO

At the heart of the Golden Agents project is ECARTICO,⁴⁶⁷ a structured biographical (prosopographical) database of Dutch artists (see Figure 5.24).⁴⁶⁸ The Golden Agents project

⁴⁶⁷ <https://ecartico.org> (Accessed 6 October 2024). See also (Brouwer and Nijboer 2017, 35): ‘Mapping the market for creative goods, both high and low, is the primary objective of ECARTICO, a comprehensive collection of structured biographical data concerning painters, engravers, printers, booksellers, gold and silversmiths and others involved in the “creative industries” of the Low Countries from circa 1475 to circa 1725.’

⁴⁶⁸ ‘The database currently contains biographical data on 65 409 persons. Painters: 9 698, Engravers: 1 437, Booksellers, printers and publishers: 3 439, Gold- and silversmiths: 7 413, Sculptors: 466.’ <https://ecartico.org> (Accessed 6 October 2024).

cross-referenced person names appearing in the ECARTICO database⁴⁶⁹ with corresponding entries in the Short-Title Catalogue Netherlands (STCN) database, the Netherlands Institute for Art History (RKDartists) database and Wikidata. The cross-referencing tasks were performed for each contributing dataset apart from ENCARTICO, which used Lenticular Lens (see Section 5.6.3).

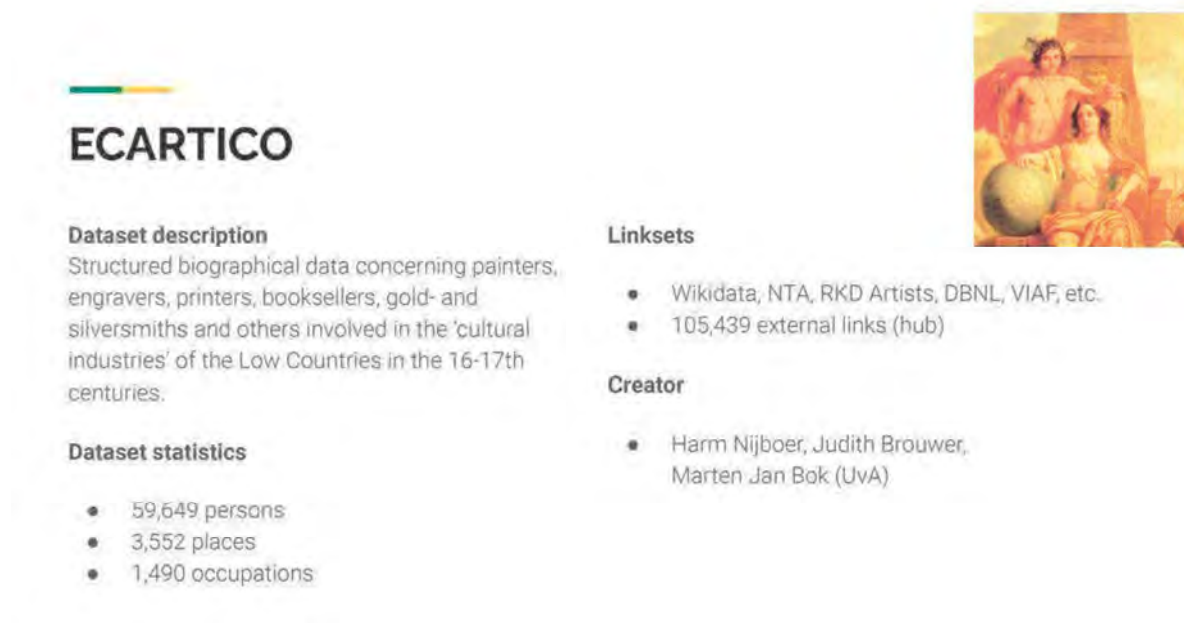


Figure 5.24 ECARTICO in the Golden Agents (<https://www.goldenagents.org/ga-output>, accessed 7 October 2024)

5.6.2.2 The Bredius Notes archive

In one part of the Golden Agents project the RKD launched a crowdsourcing project to find EBP data in the Abraham Bredius (1855–1946) Notes archive (see Figure 5.25). The project called for volunteers to attach UIs to each source and the incidence of EBP found in each

⁴⁶⁹ 'The data is (mostly manually) taken from both secondary (literature) and primary (archival) sources and is updated on an almost daily basis. ECARTICO does not only provide representations of people (and URIs) for those actively involved in creative industries, but also for their direct relatives (parents, spouses, children), customers and other relevant contacts' (Brouwer and Nijboer 2017, 36).

source. This enabled Golden Agents to incorporate prosopographical data on some 30,000 artist within the project.⁴⁷⁰

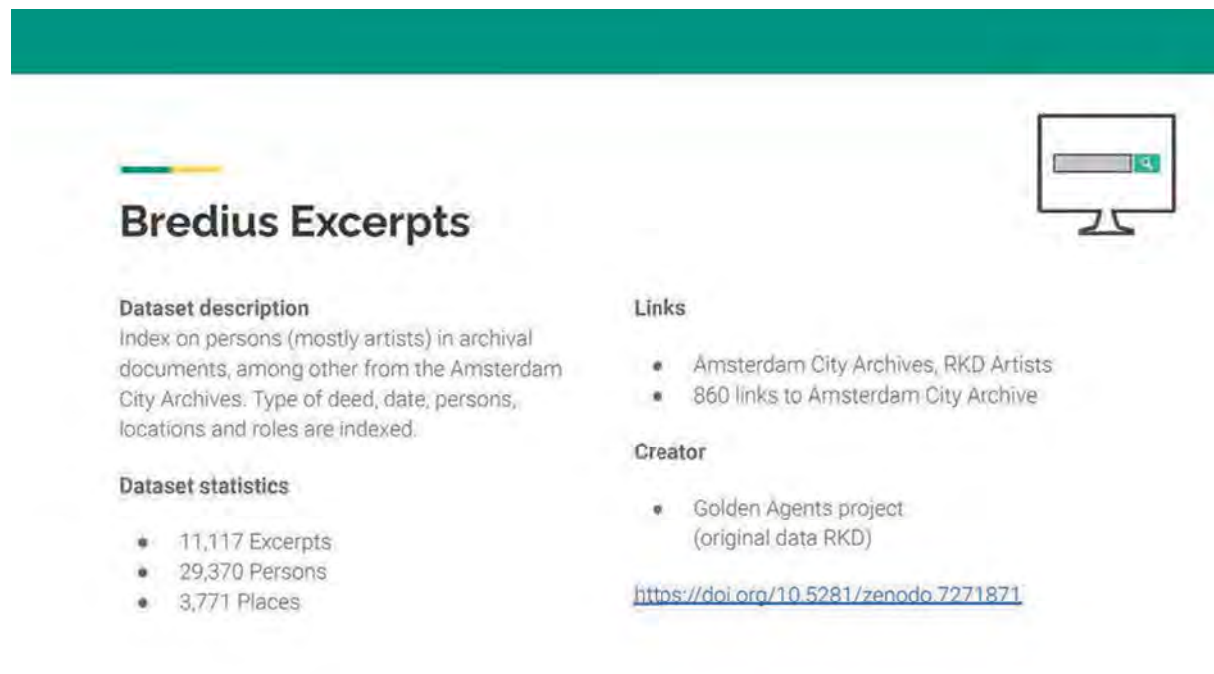


Figure 5.25 The Bredius Excerpts in Golden Agents Notary Deeds (Register of death goods)
(<https://www.goldenagents.org/ga-output>, accessed 7 October 2024)

Golden Agents also included EBP data from notary deeds records (1578–1750) held in Amsterdam City Archives (see Figure 5.26). These records include list of bequests of cultural goods (containing prosopographical data) made by citizens of Amsterdam across all sections

⁴⁷⁰ 'The goal of the crowdsourcing project is to provide the excerpts of metadata made by the archives of Bredius, so that they can be digitally searchable in the RKDexcerpts database after the project. This means that you are asked to register certain data that are mentioned in the excerpts, such as place names, dates, personal names and professions. You will also be asked to determine the type of notarial deed as much as possible.' <https://www.rkd.nl/en/about-the-rkd/job-vacancies/crowdsourcing-project-for-brediuss-notes-volunteers-needed> (Accessed 7 October 2024).

of society.⁴⁷¹ A program called TICCLAT (Reynaert, Bos, and van der Zwaan 2019) was used by volunteers to extract data which they could then link to relevant thesauri.⁴⁷²

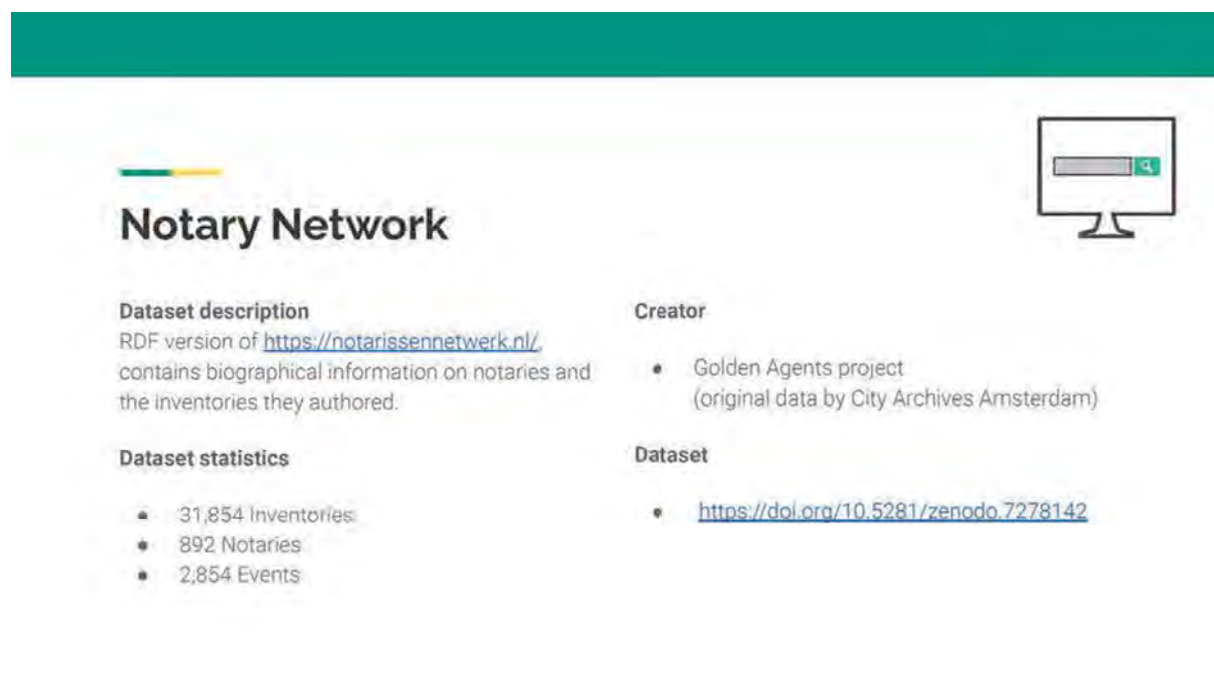


Figure 5.26 The Notary Network in Golden Agents (<https://www.goldenagents.org/ga-output>, accessed 7 October 2024)

Because each instance of prosopographical information found by the crowdsourcing initiative was allocated a UID, these were then used to compile RDF triples for the Golden Agents model, and they provided the components that enabled data integrations to take

⁴⁷¹ 'Added to this, the digitization of the enormously rich collection of the notarial deeds in the Amsterdam City Archives will provide data on the consumption of cultural goods by the inhabitants of all layers of society in Amsterdam during the Dutch Golden Age. In the Golden Agents project, novel ways are explored to extract all entities of objects that are mentioned in such notary deeds between 1578 and 1750 that are relevant to get insight into the cultural goods of Amsterdamers in the Dutch Golden Age' (van Wissen et al. 2020, 1).

⁴⁷² '[O]nce extracted and identified, almost all types of these objects can be linked to thesauri such as the Getty's Art & Architecture Thesaurus [AAT] and reconciled with textual/linguistic references to an item in an external (authored) dataset, such as the STCN, ICONCLASS, and those of the RKD' (van Wissen et al. 2020, 1).

place, disambiguation, and the establishment of a provenance trail for every data item in the Golden Agents project.⁴⁷³

5.6.3 Lenticular Lens

Lenticular Lens⁴⁷⁴ is the name of the API developed and used by the Golden Agents project to match and combine resources originating from contributing datasets using person name as the matching field (Figure 5.27).⁴⁷⁵

⁴⁷³ 'We can use these URIs to make subject-predicate-object statements following the RDF syntax. We can for instance state that the Rembrandt described in the ECARTICO database is the same as the Rembrandt in Wikidata by using the following triple of URIs: <http://www.vondel.humanities.uva.nl/ecartico/persons/6292>; http://www.w3.org/TR/owl-semantic/#owl_sameAs; <http://www.wikidata.org/entity/Q5598>' (Brouwer and Nijboer 2017, 34).

⁴⁷⁴ 'A first version of the Lenticular Lens tool (Idrissou et al., 2018) was developed by AI Idrissou at the Vrije Universiteit Amsterdam and was further developed as part of the Golden Agents in which the tool is used to interconnect resources from various heterogeneous datasets on cultural production and consumption in 17th and 18th century Amsterdam. With it, users can decide on what, how and when to link; they can cluster the matched resources, manipulate and/or validate the discovered links; they can export the links with or without their respective metadata' (Idrissou, Van Wissen, and Zamborlini 2022, 2).

⁴⁷⁵ 'Lenticular Lens, a user-friendly web interface tool that provides a set of means (data linking/integration) to an end: answering data-driven research questions (data analysis). It offers a context-dependent user-guided entity matching across multiple datasets using ad-hoc and/or off-the-shelf generic algorithms that can be logically combined. Such combination is performed over the scores of the links discovered by the various user-selected algorithms using a set of provided fuzzy logic operators. In the end, it utilizes our proposed VoID+ ontology, to intelligibly document all user-defined processes leading to the discovery, manipulation, clustering, and validation of links. All these processes can later on, at the user's convenience, be exported in various formats that include RDF and CSV' (Idrissou, Van Wissen, and Zamborlini 2022, 2).

Objective

Problem

- Research data are **dispersed** over multiple (single-scoped) datasets
- Resources inside these data are usually **not linked** between datasets
- Existing tools are **too specific** or **not flexible enough** for the specific tasks at hand in the Golden Agents project

Solution

- **Generic yet flexible tool** that works with **RDF data** in *any* vocabulary
- Tailored 'rule-based **entity linking**'
- Off-the-shelf + Tailored + Ad-hoc **matching algorithms**
- **RDF Provenance** of matched links

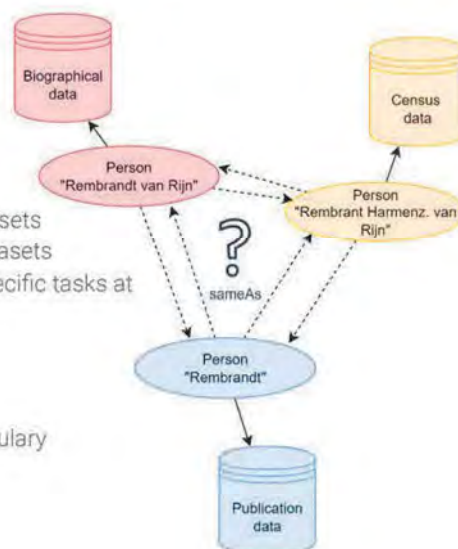


Figure 5.27 Lenticular Lens – matching names from multiple datasets (Idrissou, Van Wissen, and Zamborlini 2022)

Lenticular Lens is a comprehensive and fully flexible data matching tool. It allows partitioned data from any two component datasets to be compared for matching by using a choice of algorithms, from either those provided in the Lenticular Lens or those provided by the individual researcher, which can be matched through a rigorous process using embedded ‘fuzzy logic’. The affordance provides a matching score, enabling decisions to be made to accept or reject the disambiguation process, or re-run with new data, to match records depending on the level of veracity chosen.⁴⁷⁶

⁴⁷⁶ ‘Once explicit partition-declarations on which one wants to match entities are specific for all datasets of interest, one can now specify how entities stemmed from a partition within a source-collection are to be linked to any entity of a partition stemmed from a target-collection. For this, users are required to indicate the attributes (identity criteria) of the selected entities over which a matching algorithm of their choice should be executed. In the event that more than one algorithm is needed for various type of comparisons, the user is to specify how they are to be combined using standard or fuzzy logic operators. The combination definition is then used to compute a final identity score for each discovered link’ (Idrissou, Van Wissen, and Zamborlini 2022, 4).

5.6.4 Disambiguation

The Amsterdam City Archives (SAA) published their digitized registries as Linked Open Data (LOD). In their All Amsterdam Acts, the SAA has digitized and indexed all of the city's historical notarial acts. To resolve disambiguation uncertainties, the Golden Agents project attempted to link biographical information in the notarial acts to similar biographical data in the other datasets, achieving matches with various levels of certainty. The project recognised this was a challenging act because citizens were not uniquely identifiable, there were multiple occurrences of the same name in the records and frequently there was insufficient data to make a match.⁴⁷⁷

The project used Lenticular Lens to resolve matching disambiguation uncertainties,⁴⁷⁸ through a process of using two or more referents to accredit the same object in different source datasets (see Figure 5.28).⁴⁷⁹

⁴⁷⁷ 'This is a challenging task because (i) citizens in the Dutch Golden Age were not given any identification number; (ii) the information supplied in a single index do not suffice to uniquely identify an individual (weak identity criteria) and (iii) because of multiple occurrences of a single individual within an index.'
<https://dhistory.hypotheses.org/361> (Accessed 7 October 2024).

⁴⁷⁸ 'This process is called disambiguation and can be achieved by instantiating an identity relation such as owl:sameAs, or a more complex reified equality relation that can be qualified with information on provenance, probability, validation by human expert, and other relevant properties (Idrissou et al., 2018). Moreover, the outcome of such a process is stored in a format that allows for changes to the dataset when new information becomes available and that provides insight in the decisions taken in the disambiguation process.'
<https://dhistory.hypotheses.org/361> (Accessed 7 October 2024).

⁴⁷⁹ '[O]ur first experiments applying this tool for connecting three SAA indexes (marriage, baptism and probate inventories) and two authoritative Ecartico3 ULAN datasets' (Idrissou et al. 2018, No page numbers).

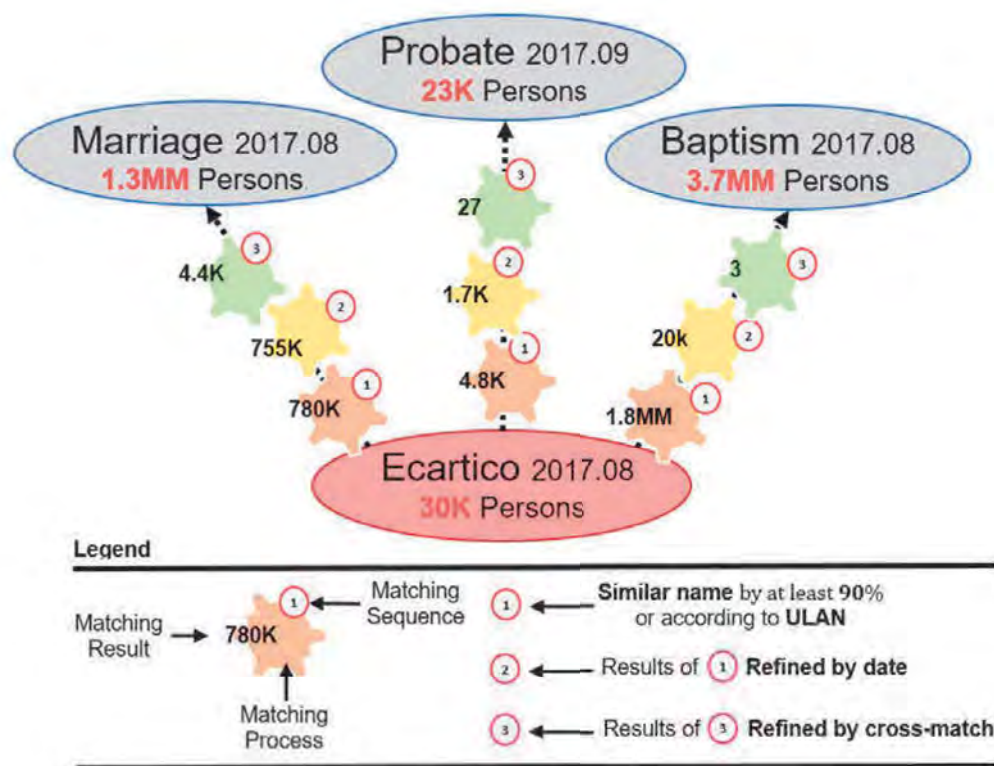



Figure 5.28 Disambiguating SAA indexes with ECARTICO (Idrissou et al. 2018)

5.6.5 Provenance

The Golden Agents project devised a ranking system to note the level of confidence they had in the determination of provenance during the disambiguation process. First, a minimum level of provenance, for items either directly observed or capable of inference from a record of a direct observation (a birth can be inferred with confidence from a baptism record, even if it cannot be proved). Secondly, a detailed provenance level based on making a scan of the source and providing a detailed annotation of all of its embedded associated prosopographical information, each element of which can then be scrutinised for matching with similar data on other datasets (see Figure 5.29).



	Min. prov.	Detailed prov.	Uncertainties
Documents / Collections	Documents grouped under certain criteria.	All the processes (creation/modification/digitization) undergone by a collection.	Who's created them? Are they trustworthy? Are they originals or copies/transcriptions?
Observation: Content	Record and its mentions/descriptions of individuals/roles and/or their types.	Annotate where the mentions/descriptions are in the text and/or scan.	Have the mentions/descriptions been modified/adapted/translated in the process? Errors introduced?
Observation: Direct Interpretation	Events and roles/objects directly observed in each document.	Break down the content into parts and connect to the observed entities and their properties	Are entities, types and properties correctly identified? Are there ambiguities? Could the observation be untrue? Uncertainty level can change.
Observation: Indirect Interpretation	Events and roles/objects that can be indirectly inferred.	Which extra events, individuals and properties can be extracted from the ones already extracted or from the original content.	What type of inferencing? Probabilities involved.
Reconstruction	Several observations can be combined to an individual concept through time.	Which specific properties were used for disambiguation.	Is the provided information enough for disambiguation? How accurate is the disambiguation process?

Leoni van Wissen, Veruska Zamborlini, Challenge LODLAM 2020, Los Angeles, Getty Research Institute, February 3th-4th, 2020 <<https://lodlam.net/challenge-entries/>>

Figure 5.29 Modules with levels of detail and uncertainties (van Wissen, Zamborlini, and van den Heuvel 2022)

The Golden Agents' thorough and comprehensive data matching and disambiguation process provided a granular and detailed approach to extracting and verifying prosopographical data across all of the data of the Golden Agents project. Figures 5.30–5.32 show how provenance is built up over a succession of layers, with each observation recorded and identified with its own UID.

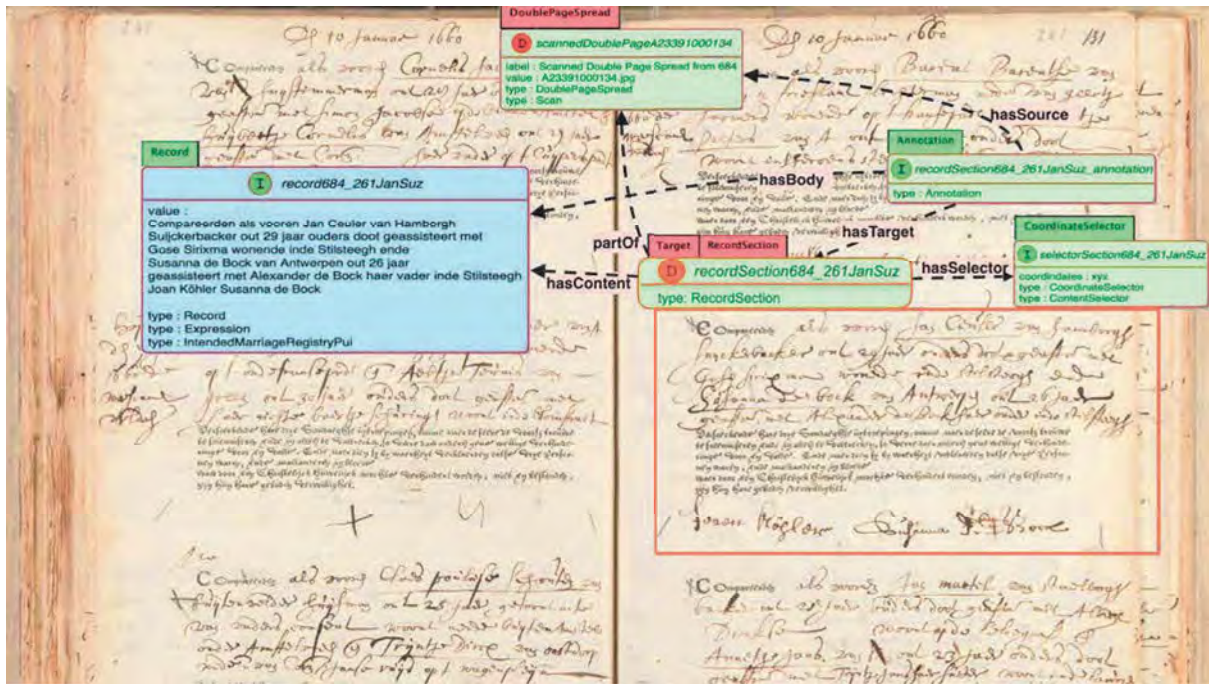


Figure 5.30 Building provenance in Golden Agents: Step 1 – locating and identifying data

(<https://www.goldenagents.org/ga-output>, accessed 7 October 2024)

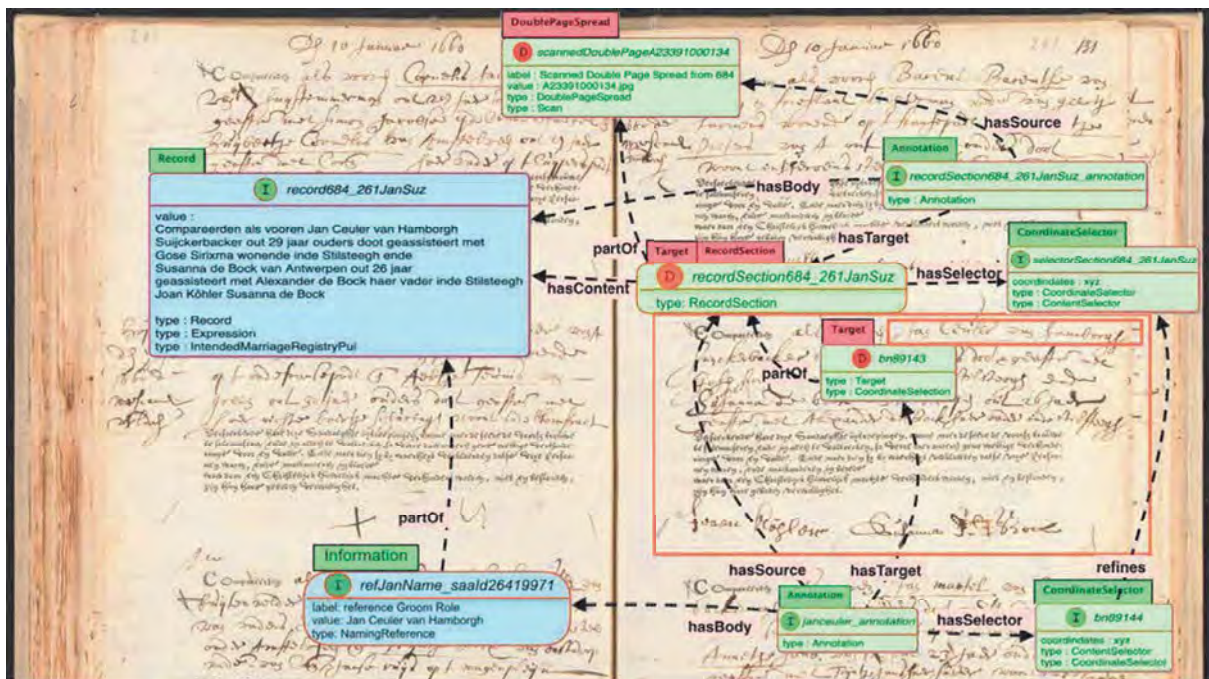


Figure 5.31 Building provenance in Golden Agents: Step 2 – linking up marriage data

(<https://www.goldenagents.org/ga-output>, accessed 7 October 2024)

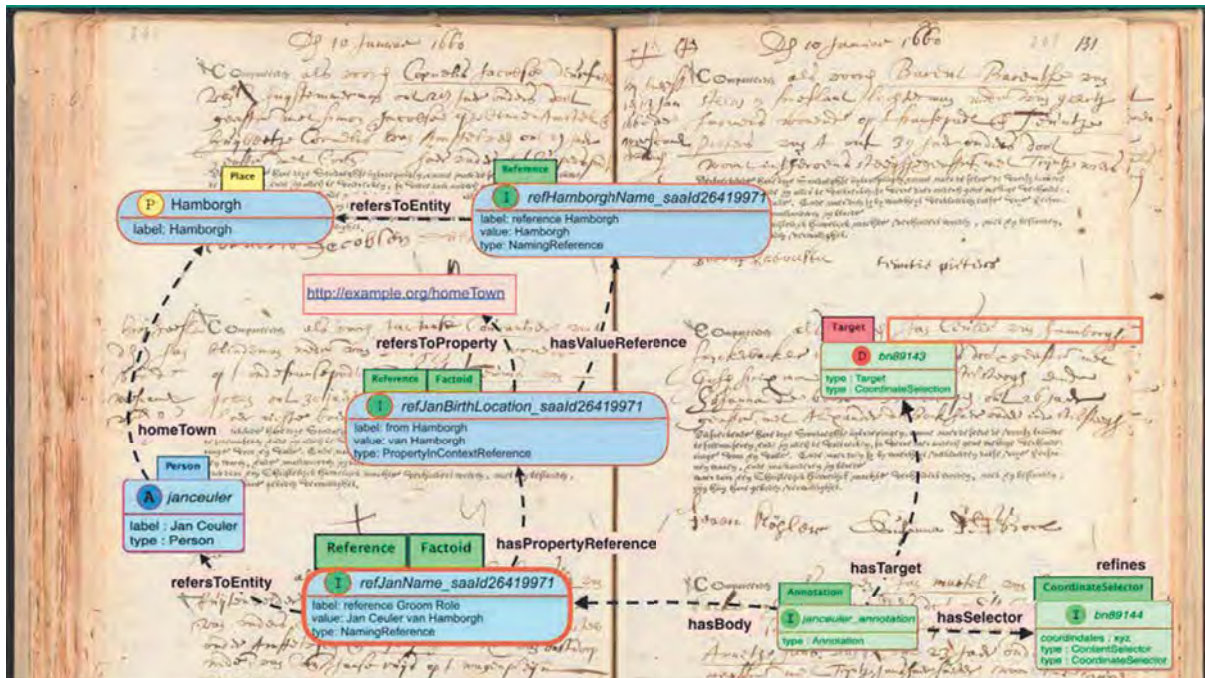


Figure 5.32 Building provenance in Golden Agents: Step 3 – linking up locational data (<https://www.goldenagents.org/ga-output>, accessed 7 October 2024)

5.6.6 Browser

The Golden Agents Dataset Browser is a graphical interface built on the Timbuctoo platform.

It allows the user to get a clear overview of the combined dataset, its metadata and its

classes and properties. The entire dataset can be searched using SPARQL (see Figure 5.33).

levels. The scope for an NAI-UID system to enhance the Golden Agents project was discussed favourably with Professor Charles van den Heuvel in September 2024. The Golden Agents project is clear evidence that when used in LOD schemes, EBP provides robust provenance, offers a structure in which to manage a rigorous disambiguation process and affords full compliance with FAIR principles. It facilitates a structured, flexible and open system for the linking up of data across multiple archives using prosopographical information at both access and joining points.

5.7 Research projects review

Project Title	Social Networks and Archival Context	Cambridge Group Integrated Census Micro-data	Traces through Time	ResearchSpace	Golden Agents
Project infrastructure regime (Section 5.7.1)	US	UK	UK	UK	EU
Archivist/researcher balance (Section 5.7.2)	Both	Research	Archive	Archive	Both

Evidence Based Prosopography (Section 5.7.3)	Mixed	Yes	Mixed	Mixed	Yes
Data management (Section 5.7.4)	Open	Closed	Closed	Closed	Closed
Genealogical data component (Section 5.7.5)	No	Yes	No	No	Yes
Provenance (Section 5.7.6)	Database	Primary source	Database	Database	Primary source
Technologies and integrated tools (Section 5.7.7)	Yes	Yes	No	Yes	Yes
Usability and user interventions (Section 5.7.8)	No	No	No	Yes	Yes
Visualisation affordance (Section 5.7.9)	Yes	No	No	Yes	Yes
On-going support (Section 5.7.10)	No	No	No	No	No

Table 5.3 Chapter 5 project review table

Table 5.3 presents a projects review comprising ten project/data descriptors. Firstly, the home region of each project is discussed to discover if regional characteristics that shape and define the nature of the projects are evident and whether each of the project presents an affordance primarily for archivists, other researchers or both.

5.7.1 Project infrastructure

SNAC is based in the US with a mix of academic and philanthropic sponsorships that developed over time and as the project was built up and grew to maturity. Project participants, similarly, grew over time. SNAC began small and evolved as interest in the project and the number of project participants grew. SNAC is an example of a ‘bottom-up’ project, championed by one individual and with different strands of project support, gaining gradual buy-in by financing and nurturing institutions which fell into place as the project developed. Over time the project grew and the number of project participants gradually became active participants.

At the other end of the spectrum, the Golden Agents project in the EU benefited from a ‘top-down’ infrastructural design throughout. The Golden Agents project lead,⁴⁸¹ the Huygens Institute for the History of the Netherlands, designed the project together with

⁴⁸¹ Meeting with Professor Van den Heuvel, project lead, 24 September 2024.

twenty co-applicants (other Dutch DH institutions).⁴⁸² The project was supported at its inception by six collaborative partners and included a twelve-person steering group. Additionally, the project had a five-person scientific advisory board and a four-person technical advisory board. This was a CLARIAH project and it was funded by CLARIAH throughout. The project built on fourteen previous EU-supported DH projects including ECARTICO, the Bredius Notes Archive and national digitised Notary Deeds. The project used dedicated software and technical expertise from a centralised dedicated DH research support hub, LAB1100. The project unified and interconnected with a wide range of recent EBP based digital projects in the Netherlands, bringing together considerable relevant expertise and data knowledge from a wide range of Dutch DH agencies to support and deliver the project.

The three UK based projects (Cambridge Group Integrated Census Micro-data, Traces through Time and ResearchSpace) all began as single research hub projects, each of which reached out and formed links with other research hubs and data providers as they developed. The Cambridge Group had a necessary commercial partner in FindMyPast (the owners of the project data), but FindMyPast was not an active contributor to the project; its role was to restrict the project's use of data to protect its commercial interests.

ResearchSpace was supported financially throughout by grants from the Mellon Foundation. The Traces through Time project was funded by the Arts and Humanities Research Council.

The UK projects illustrate the hybrid nature of UK infrastructure. Smaller than either the US

⁴⁸² 'The Huygens Institute is an institute of the Royal Netherlands Academy of Arts and Sciences. This Akademie is funded by the Dutch government. In addition, a significant part of our Institute's research is funded on a project basis by the Netherlands Organisation for Scientific Research (NWO), by the European Union (Horizon Europe), and sometimes also by private funds, foundations, heritage institutions and local governments.' <https://www.huygens.knaw.nl/en/informatie-2/about-huygens-institute/mission-vision-and-strategy> (Accessed 7 October 2024).

or the EU, they exhibit a range of project infrastructure designs using both US and UK models adopted variously by each project.

A consideration of these infrastructural features and characteristics suggests that establishing EBP as the bedrock structure of all EBPD, although a universal attribute, might best be approached differently in the US, the EU and the UK. In the US it can be anticipated that an approach to the research philanthropic foundations together with commercial genealogy companies might be appropriate. In the EU, an approach can be made directly to the community infrastructure agencies, EADH, DARIAH, CLARIAH and CLARIN, because financing and project support can be efficiently organised centrally. In the UK, an approach to the established humanities research hubs based in universities, together with a collaborative approach to the GRO and FindMyPast,⁴⁸³ might provide effective forums in which to pursue EBP and the mapping of EBPD.

5.7.2 Archivist/researcher balance

The DH projects considered here are all targeted towards GLAMS archivists who use digital tools to manage their catalogues, finding aids, indexes and collection descriptors. They are also directed towards general researchers who might use digital tools to locate and work with data. However, the extent to which each project actually offers support to these two communities varies. Only one, The Cambridge Group, offers a data service directed solely at general researchers. This perhaps explains why all of the data gathered and re-presented for researcher use is presented only in its cleaned and organised form, and project output data

⁴⁸³ WikiTree is a suitable open-source alternative genealogical provider.

is made available as one single data service (homed at UKDS). The Cambridge Group project includes a full, detailed and comprehensive user manual to assist data users, with over 2750 data users registered at the data store in 2020 and 2021.

TTT and ResearchSpace have encountered considerable difficulty as general researcher affordances, in both cases due to the high level of complexity of the project as a whole and in terms of the considerable data manipulation undertaken and mixed technological skills assumed to be present in the general researcher community. It is for this reason that the TTT project has stalled as a public offering, though ResearchSpace continues to be used as an in-house tool at TNA. Both the SNAC and Golden Agents projects strike a balance between archivist and general researcher users, but only the Golden Agents project points to a long list of research projects undertaken using its platform.⁴⁸⁴ Golden Agents has integrated well within the Netherlands and wider EU archivist community (largely due to its structure by design; see Section 5.7.1). SNAC does not report similar information in this regard, but an exhaustive search for such projects has not revealed other research projects that have majored on using the SNAC service. SNAC does however have clear indications of a high level of take-up in the archivist community, evidenced by the registered users of the system and the number, quality and recorded attendees at SNAC training events.

5.7.3 Evidence Based Prosopography (metadata versus primary source)

Only two projects were based exclusively on EBP. In respect of the Golden Agents project, all of the EBP elements collected by the project from a variety of EBPD Records were direct

⁴⁸⁴ Via the project website <https://www.goldenagents.org/ga-output> (Accessed 29 November 2024).

refers to a located and identified prosopographical primary source. In respect of The Cambridge Group project, again, all of the data collected by the project was directly related to prosopographical information – the nineteenth-century UK population census data.

The other projects collected only GLAMS catalogue data which may in part consist of embedded references to prosopographical information, but the project datasets do not (and cannot) make a distinction between data elements that are representations of information in Records and those that are derivative GLAMS catalogue data. Therefore, data provenance for all three projects resolves only to the GLAMS catalogue data itself, and no further. For the general researcher to rely on data present in these projects is to accept GLAMS catalogue data as the proxy data source. There is therefore, for these projects, a disconnect from the Records themselves, which can only be resolved by general researchers' detailed critical examinations of each individual data element by finding the primary source from which the metadata reference may have derived.

It is encouraging for this thesis that two of the projects examined here (The Cambridge Group and Golden Agents) have chosen to collect, manage and present only EBP and for each project rigorous data provenance management is a key feature. This thesis urges that clear and identifiable separation should be made between GLAMS metadata, referential data and Representative Data (with provenance tracking back to Records).

5.7.4 Data management

ResearchSpace is designed to allow users to add multiple metadata sets to the system as long as they conform to universal data specifications (see Chapter 4) and are capable of

being handled by the ResearchSpace system. Users report that the system requires a high level of technical expertise, which discourages use of the system by all but a small group of highly trained individuals. SNAC accumulates data (comprising solely metadata) only through its system of participating data donors. metadata follows a prescribed content and formatting system based the underlying metadata structure itself, and it is effected through the design of the project data extraction and data homogenisation algorithms. Users cannot independently add or remove datasets from the system. In the SNAC project data is protected, restricting users to the datasets and the SNAC system functionality. This 'closed' data system is a trade-off against the risks of data loss, corruption and misuse that 'open' systems can invite, and the risk is managed by preferring future user engagement only from highly skilled users (archivists). TTT intends to be a closed system preventing users from adding, removing or changing underlying data, allowing users to view relationships between data elements that the system generates. Golden Agents is also a closed (or arguably mixed) data system because it is designed to be open at a later date, when new centrally organised research projects seek to reuse, modify or adapt the project datasets for other uses. The Cambridge Group is the most closed system, not allowing users to interact directly with project data at all.

Data management complexity due to the complications involved in merging data from multiple datasets is one of two major concerns for all projects (the other is technical complexity – see Section 5.7.8). Each of the project's literature talks at length about data incompatibility and data complexity and the considerable efforts all of those involved made to overcome these challenges. The solution each adopted is bespoke and results in archivists and general users needing to have a deep appreciation of both the data challenges met and the complexity of the consequential solutions. This disadvantages the

general researchers compared to the archivists. Later users of either type, if unfamiliar with the technological aspects of the projects, can find these challenges and their respective solutions difficult to fully comprehend. This then can lead users to experience data provenance and data type concerns (is the output derived from metadata, referential data or Representative Data?) that are not easily overcome by those without relatively high levels of both data and technological handling skills.

5.7.5 Genealogy

The Golden Agents project, because it is linked to the ECARTICO database, including its genealogical data structure, facilitates data selection and data analysis using genealogical data. The Cambridge Group project is totally composed of genealogical data. Its functionality, however, is restricted because its commercial arrangements with FindMyPast (the data provider) do not allow FindMyPast genealogical analysis functionality to be enabled in the publicly available version of the project software. This restriction can be overcome by the acquisition of general user project-specific licences from the data owner. The other projects do not provide genealogical data for analysis because GLAMS data does not specifically include genealogical data, though it may from time to time appear unstructured and often unattributed in GLAMS data Notes Fields. It is one of the key objects of this study to encourage the open adoption of both genealogical data and the related NAI. It is encouraging that two of the projects analysed here embrace genealogical data as a key data component and adopt the NAI-UID elements of a source and a project UID.

5.7.6 Provenance

Two of the projects, The Cambridge Group and Golden Agents, have gone to great lengths to establish data provenance tracking back to the Records; the others rely solely on the metadata provenance regimes of the various data contributors to establish data provenance. Donor metadata provenance tracking can (and often does) vary from one metadata set to another, as Chapter 4 has set out.

The analysis in Section 5.2 has indicated that the SNAC project may be more useful as a tool for archivists (who well understand the strengths and weaknesses of their own datasets), but is of questionable utility to the general researcher, who is often unfamiliar with the complex history of metadata standards and systems. Golden Agents also provides a research resource and enquiry tools to the general researcher, but by the project's careful selection of only EBPD, and the efforts it makes to establish provenance back to Records, provenance concerns are largely overcome. The Golden Agents project deliberately strove to ensure its usefulness to the general researcher, evidenced by the considerable efforts that it made to engage with general researchers. (This is also evidenced by the high level of general researcher take-up – see Section 5.7.2.) The Cambridge Group project also offers data with a demonstrably high level of clarity for the general user, which derives from the project's discipline in respect of data management, provenance tracking back to primary sources and the availability of a comprehensive user manual. TTT and ResearchSpace are both projects aimed predominantly at the archivist community and less so the general researcher. Both feature only GLAMS metadata and so leave provenance issues to be managed by each of the

contributing data users. TTT project development has stalled until person name provenance issues can be satisfactorily resolved.⁴⁸⁵

Provenance is a key concern across DH for both archivists and general researchers.

Archivists, through training, skills and deep familiarity with their metadata systems, have an institutional advantage over the general researcher, who often lacks both the necessary technical skill to engage critically with the model and deep familiarity with the complexities of metadata.

Across these five combined (meta)datasets, provenance rules methodologies and practices are project specific and may not be equitable. This brings into serious question the future use of complex combined datasets, which are unlikely to be completely interoperable one project dataset to another. This is particularly so if later research projects use data (in whole or in part) from these five projects and then merge that data with other data taken from other datasets (perhaps newly made from the researcher's own data collecting activities) to form yet more combined datasets, which are then in turn offered through data stores to yet more researchers. Without addressing the importance of the wider world of data provenance immediately, the future of datasets looks problematic.

This project has argued for the adoption of an NAI to help in part to manage data diffusion and data uncertainty, at least at the national level. This depends on the separation of metadata, referential data, referent data and primary data. A first step is to structure and systematise the EBP to establish a system of EBP data linking and provenance tracking to help manage the growing amount of poorly organised digitised information on PHL.

⁴⁸⁵ Meeting with Mark Bell, Senior programmer at TNA 27/10/2023.

5.7.7 Integrated tools

SNAC, ResearchSpace and Golden Agents built integrated tools for users to interrogate and use the system. TTT similarly intended to build a user tool, but it is not yet certain whether, when or how that might be completed. The Cambridge Group built integrated tools to allow the project team to achieve its goal, the provision of homogenised PHL data spanning several national censuses, but did not build a user tool. All of the projects found it either necessary or desirable to build integrated tools both for the project team to achieve project goals and as an affordance to help general researchers to engage later with the project data, and each project evidences varying levels of success. Nevertheless, integrated tools were necessary for the projects to achieve their objectives:

- To enable all data to be held in one project container.
- To facilitate the transition of data through the system from source to destination.
- To allow tracking of data throughout the system.
- To facilitate data auditing (avoid data losses and duplications).

How to develop integrated tools for user engagement was an observed point of difficulty for all of the projects, due to the anticipated skill level gap assumed to be present among users. SNAC overcame this difficulty because its user community was chosen to be almost exclusively made up of archivists who were familiar with the data. To support that, extensive customised training on using the SNAC system has been provided.⁴⁸⁶

⁴⁸⁶ '[U]sers from academic domains tended to be more persistent and use different search strategies to reach their goals. This is important, since it suggests that academics are keen to find and potentially to use certain

ResearchSpace is a user tool allowing researchers to access a large array of data sources and then construct research assemblages using powerful relational and visualisation technologies on a single research surface. The development team has used ResearchSpace in a small number of commissioned projects. Several new researchers have successfully used the Golden Agents tool set. However, ResearchSpace and Golden Agents face difficulty in future user take-up due to the technological and data complexity of the projects, a problem that will be exacerbated as project teams fall away and new users struggle with their complexity.⁴⁸⁷ TTT has yet to respond to the technical complexity dilemma and The Cambridge Group has shielded users from exposure to technical complexity by focusing on the data itself as the project deliverable.

5.7.8 User interventions

User interventions fall into three broad categories: the abilities to (1) change the underlying data, (2) add or remove data in a custom application and (3) shape the data output.

SNAC invites users to change the underlying data through a clearly defined, managed and audited change and approvals process. None of the others allows underlying data to be changed. The Cambridge Group, because it does not offer the data processing system itself to other users, does not offer data manipulation under any category.

resources. Nevertheless, we would argue that the kind of scholar who is likely to know they need such a resource and persist until they find it is the kind of early adopter who is already using specialist digital resources' (C. Warwick et al. 2007, 100).

⁴⁸⁷ '[T]oo many digital resources require users either to struggle with unfriendly interfaces or to be technical experts even to begin to use them' (C. Warwick et al. 2007, 100). See also (Terras et al. 2018).

The remaining three projects all offer data manipulation under categories 2 and 3. The concern is that these three projects are technology and know-how rich. ResearchSpace has already come under criticism for its complex technicality. TTT has not yet fulfilled its objectives and so a user response cannot be ascertained. Golden Agents has had considerable user success, evidenced by the several research projects completed since it went live. Future use when the project team disperses is uncertain.

There is no doubt that all of the four projects offering some level of user intervention are both technology rich and technology complex. They all rely heavily on the latest technologies and their conditional techniques, and these may soon become out of date.

5.7.9 Visualisations

The deployment of data visualisation technologies is an indicator of projects using, where possible, the latest technologies. Once again, The Cambridge Group is excused in this category because it did not make a user tool. But interestingly, neither did The Cambridge Group incorporate visualisation technologies into the mix of technologies used by the project team itself. It could have done so, as the Golden Agents project shows, where using visualisation technologies to analyse data management processes was shown to be very successful. Golden Agents has prioritised the use of visualisation techniques in the area of data management and less so in the area of data output, leaving new users to choose (or not) to use the project's data user technology, Lenticular Lens. ResearchSpace's key deliverable is a workspace on which selected elements from a variety of data records can be visually displayed and elements then spatially arranged in a relational format.

Visualisation technologies thrive in the current drive to develop the Semantic Web, and they attract considerable attention from users and funding agencies. That they can be successfully deployed both by the project teams during project build and also later by archivists and general researchers illustrates the way in which visualisation technologies have been variously adopted by these projects and become an integral part of the project as a whole.

5.7.10 Ongoing support

The Cambridge Group project, having presented its output in the form of a unified national census dataset, would appear not to require ongoing support. However, the considerable extent to which the project used data management processes and data handling techniques to achieve its aims means that future users rely on the project's manual for detailed explanation of the processes and techniques used and how they have impacted the data deliverables. The Cambridge Group project team is now dispersed and the project website indicates that further support from the team is no longer available. This concern, detachment of the project data from the project team, is in part mitigated because the project team did not adopt a complex technological solution (for example, no visualisation technology dependencies).

SNAC enjoys the highest level of ongoing project support through its user groups and the familiarity users have gained from using the system on a daily basis. These indicate that future use of the project's deliverables among archivists is likely to be well supported.

Reliance on the original project team is therefore low and in the future, changing or replacing the system (on the face of it) should not be a burdensome task.

It seems likely that both ResearchSpace and Golden Agents are, in some sense, fixed in time and that in the future it may well be the case that these projects are abandoned but the data generated saved. In this sense they will have probably served their purpose, and each perhaps at the outset assumed that technologies in this area are changing and are likely to continue to change rapidly. For SNAC, ResearchSpace and Golden Agents the technology is the main deliverable; for TTT and The Cambridge Group the data is more important than the technologies.

5.8 Durability lessons from the five research offerings studied

This chapter has summarised a view of each project's likely durability based on the project table and its analysis (Table 5.3).

In the case of SNAC, the project is highly durable:

- The intended users of the tool developed the project infrastructure responsively and organically.
- It made provision for general users, but the project's intent from the outset was to be a daily working tool for the user group of archivists.

Because the project's source data was archivists' metadata, the project was not troubled by a need independently to verify the accuracy of any embedded EBP. This suggests that the tool would be of less use to the general researcher, who might well need greater use of primary, verifiable data for research.

- The data management scheme is a live, continuous scheme performed by the archivist members themselves. It is therefore highly durable.
- The project does not use genealogical techniques but instead uses a 'relationship' tool based on the number of appearances of a data element (person name) in the integrated metadata set. The system then regards the presence of any two or more names in any one metadata set as suggesting a relationship between the two persons. While this visual representation is of some research value, this is perhaps too crude a relationship test to be of considerable value to the general researcher (even in the spirit of the Semantic Web).
- Because the system does not allow data manipulation and it is used almost exclusively by archivists, provenance lies with the owners of the contributing metadata sets, not burdening the SNAC system with too much complexity.
- Because of the close relationship between metadata owners and metadata users and the high level of engagement with the system by users, the project appears therefore to be highly durable.

For The Cambridge Group:

- The project had a specific start and end date, and the project team is now dispersed. The project tool has fallen away and only the manual remains as a detailed explanation of the project itself. The EBPD the project produced has proven to be highly valued by many other users, who would have needed otherwise to reproduce and re-run the model with the same data. It is not certain that FindMyPast would contract for later related projects similarly to use data which is contractually reserved for its own commercial use.

For TTT:

- This project has not yet completed, and completion seems to depend on a reliable means of matching Person Names among the metadata about the considerable and broad collections held by TNA. The project would benefit from the application of the NAI to allow instances of person names to be tagged permissively to an NAI-UID index entry (an EBP representative source).

For ResearchSpace:

- The project is an in-house tool deployed at TNA and the British Museum. It is complex by design and difficult to use without considerable training. It relies heavily on integrated technologies that are likely to be rapidly made redundant. The project is a groundbreaking affordance demonstrating the capabilities of digital technologies in the humanities. Its low level of durability does not detract from the way in which the project has demonstrated what is possible using current and anticipated new relational technologies.

For Golden Agents:

- In its deployment of technologies and data handling affordances, Golden Agents perhaps shares a similar future durability path with ResearchSpace. But this project is highly integrated with other substantial humanities projects in the Netherlands and therefore its future success is bound up with that of others. In this writer's view the greater success of the project is its exclusive use of EBP and the development of effective and proven provenance and data management systems that function across multiple datasets. The technologies may well fall away over time, but the

project's contribution is the practical application of information science theories and techniques to the problematic and demanding field of using EBP in DH.

The main focus, and interest, in all of these projects (except The Cambridge Group) has been on emergent technologies, but what is revealed through this detailed study is the concern they all have for data, whether metadata, referential or Representative Data. These projects demonstrate the considerable time and effort given to the care for data, and therefore the respect given to data, irrespective of the technologies each project chose to use.⁴⁸⁸ They illustrate that it has been right for this thesis to choose to examine carefully and critically the world of data on PHL rather than technologies, and to seek to find remedies for the common desire across the humanities to be able to explore data as evidences of PHL, systematically, rigorously and confidently. This can only be achieved if the current attention given to technologies in DH is matched with a similar disciplined approach to the data and the information sciences these tools exploit. Robert Crease, Elyse Graham and Jamie Folsom recognise the need to rethink the nature of past data in an age rich in data technologies,⁴⁸⁹ but it is EBP that points the way ahead for DH and the other academic disciplines that seek to access data on PHL.

⁴⁸⁸ '[H]umanities users are highly critical of the quality of research resources themselves, thus content must not only be of excellent quality, but must advertise this fact, by making clear what kind of material it contains and how this has been selected' (C. Warwick et al. 2007, 100).

⁴⁸⁹ 'In this era of big data, huge international collaborations, less visible scientific activities, and infrastructure, it is more critical than ever to create new models and tools to support a new kind of history of science. If we wish for historians to explore large-scale science in a way that they have not been able to do before; for the general public to easily explore openly available government data in order to improve the public understanding of science; and for government officials to better understand what happens to the money spent on individual research projects, then we need to find ways to consolidate the history of such facilities and make the data of that history available for scholars and other users to work with more easily. Without doubt, the triumph of invisible knowledge infrastructures—the deep structure of databases rather than the visible structure of card catalogues and library classification systems—will dramatically change our methods of knowledge management in the humanities. But we will not have lost the need for humanistic approaches to knowledge management until we have ceased to reason using analogy, narrative, and metaphor' (Crease, Graham, and Folsom 2019, 55).

Chapter 6 EBP in Project Seven

To demonstrate how EBP can be used in practice in DH research projects I undertook a research project working as a Independent Researcher. The project was named Project Seven (P7) and a Project Seven Report was generated from it (Appendix 7). The P7 report will be offered to the Quaker Studies Research Association as its first fully digital project on completion. To answer the question: how successfully has the small-scale, unfunded, P7 research project into Past Human Lives used Evidence Based Prosopography?, the P7 report is explained in detail here.

I have found one study which is a close comparator to Project Seven, John Haggerty and Sheryllynne Haggerty 2011, *The life cycle of a metropolitan business network: Liverpool 1750–1810* (J. Haggerty and Haggerty 2011). Haggerty and Haggerty say, ‘This paper pushes forward this more nuanced and sophisticated analysis of networks and represents the first serious attempt to measure them to assess change over time’ and that the study, ‘uses visual analytics of Liverpool's business networks comprising political, trade, social and cultural institutions to assess their [mercantile networks] role in the changing social and economic climate during the period 1750–1810’ (J. Haggerty and Haggerty 2011, 189).

Haggerty and Haggerty’s network analysis project consists of a cohort of 210,000 relationships between 1700 actors, active between 1750-1810 using SocNetV software.⁴⁹⁰ In comparison, Project Seven is a network analysis of some 12000 relationships between 3600 actors, active between 1830-1870. Haggerty and Haggerty have modelled society memberships between 1700 actors and 5 social groups (Liverpool Town Council, the

⁴⁹⁰ <https://socnetv.org/> (Accessed 17/04/2025)

Committee of the African Merchants Trading from Liverpool, the Library/Lyceum, the Mock Corporation of Sephton Club and the Ugly Face Club). Their study assumed that, ‘all members listed within one decade met each other’ (J. Haggerty and Haggerty 2011, 193). Perhaps the number of relationships was also counted once for each year of membership per person, (the study cuts data into decadal groupings),⁴⁹¹ these two factors may account for the seemingly large number of relationships in the Haggerty and Haggerty study.

Haggerty and Haggerty were interested in power relationships and especially in the commercially sensitive commissioning of international trading ventures, using ships from the port of Liverpool, (often slave trading).⁴⁹² Their previous considerable research into the mercantile past of Liverpool led them to consider the social groups studied to be possible locations of commercial dealings and power brokerage amongst Liverpool’s maritime traders, therefore the memberships of the social groups and the movements of individuals between groups, could be relatable to trade conditions and activities which they had previously studied (S. Haggerty 2006), (S. Haggerty 2008). The interpretations Haggerty and Haggerty make in their network analysis relies on their prior knowledge of the contemporaneous Liverpool trading communities. Haggerty and Haggerty recognise that a network study alone (without complimentary archival research) can ask more questions than it answers.⁴⁹³ Haggerty and Haggerty’s project, as described in *Explorations in*

⁴⁹¹ The network analysis of institutional membership is broken down into three groups of two decades: 1750–1769, 1770–1789 and 1790–1809. (195)

⁴⁹² We have used the Trans-Atlantic Slave Trade Database (Eltis et al., 2010) to follow investment trends in this trade by actors in our networks. Whilst we recognise that this trade was only one of many interests that these actors will have had during this period, this source provides evidence as to how Liverpool’s merchants used their institutional membership networks to engage in economic activity. In addition, it supports our argument that, far from being static, networks and the way that actors used them was dynamic. (195)

⁴⁹³ Using visual analytics has facilitated the analysis of a large data set with over 210,000 relationships involving 1700 actors. Moreover, it has provided a nuanced and sophisticated view of the way in which institutional membership networks changed over time, not only in the long term, but dynamically within the short term as well. This study confirms that using tools such as visual analytics can be useful in raising both

Economic History 48 (2011) 189-206, is comparable to Project Seven in terms of its data, methodology and technologies. It is also a 'lone' researcher model project and illustrates that the approach taken here in P7 mirrors at least that taken in one other research project.

This chapter proceeds in the style of Chapter 5. The P7 project and the P7 Report are examined in detail and scrutinised in the way that other recent research projects are scrutinised in Chapter 5. This chapter ends with the P7 project considered within the critical analysis framework used in Section 5.7. The major difference between the P7 project and those considered in Chapter 5 is that P7 is a much smaller project – conceived and executed by a 'Independent Researcher' (the thesis author), unfunded and minimally supported, whereas the projects assessed in Chapter 5 are much bigger, involving large teams with wide expertise and with considerable resources (including funds). A striking observation is that P7 does not stand out as a lesser project in comparison to those discussed in Chapter 5. First the P7 objectives are set out (Section 6.1). Then the origins and background of the project are introduced, including a timeline from project inception to implementation (Section 6.2). Thomas Hodgkin MD is introduced and his central role explained: through his social networks he brought together 600 Quakers and then 2400 others (the Quaker Led Group, QLG) who were initially concerned to relieve the plight of aborigines throughout the British colonies, and who then turned to institution building in anthropology (Section 6.3). In Section 6.4 how Quakers influenced the wider QLG is described. Section 6.5 discusses the issues that arose (and were addressed) in establishing and qualifying the QLG membership for the purpose of this study (who is and who is not included).

specific and wider issues from the quantitative data, especially where there is a lack of extant qualitative data. In addition, visual analytics is exploratory in nature, and therefore may raise questions, rather than answer them per se. (204)

The work of the QLG spanned several societies and these are set out in Section 6.6. I collectively call them the Centres for the Emergence of the Disciplinisation of Anthropology (CEDA) in Britain 1830–1870. Section 6.7 discusses the first of these societies, the Quaker Committee on the Aborigines. It is from here that Hodgkin began his lifelong work to promote the welfare of aborigines throughout the British colonies, and his passion for his scientific interest in anthropology. Section 6.8 discusses the parallel House of Commons Select Committee on the Aborigines and the support structures for the work of the Committee – Hodgkin’s fellow Quakers in Norwich. Section 6.9 discusses the Aborigines Protection Society (APS), which Hodgkin formed at Ratcliff Meeting House in June 1837. Section 6.10 describes the Ethnological Society of London formed out of the APS by Hodgkin and a few colleagues. How these societies became the birthing societies for the discipline of anthropology in Britain is set out in Section 6.11.

Section 6.12 describes how P7 identified and modelled the significant presence of extended Quaker families among the 600 Quakers in Hodgkin’s networks. Sections 6.13–6.17 detail the building of the Human Data Digital Toolkit (HDDT), the suite of technologies used in the study, and Section 6.18 outlines how the HDDT was populated with data taken from a range of data providers and in a variety of formats. Sections 6.19–6.22 discuss the three Case Studies performed on the model data. Finally, in Sections 6.23 and 6.24 P7 is reviewed and the project conclusions are drawn.

6.1 Project Seven objectives

The aim of P7 was to explore the relationships between Thomas Hodgkin MD (1798–1866) and 600 Quakers among 3000 activists who came together to relieve the plight of aborigines

throughout the British colonies and, acting out of that concern, from 1830 to 1870 the societies that would lead to the formation of the discipline of anthropology in Britain (the CEDA). The first third of P7 research time was the acquisition of data; the second third was the design and build of a SQL database and associated technologies, the HDDT.⁴⁹⁴ The final third of the P7 time was using the database and the data visualisation technology to explore the dataset and construct three Case Studies. The Case Studies were performed in Jupyter Notebooks (JNB) using Gephi data visualisation software. JupyterBook (JB) was used to produce the Project Seven Report and it is reproduced at Appendix 7. This chapter makes several references to that report. The P7 project is also online at: <https://kelvinbeerjones.github.io/project-seven-book/intro.html>. (The images reproduced here are low-quality versions of the P7 project images which can be viewed at full resolution online.)

Expanding on the lines of enquiry described above, I established three questions that analysis of the data in P7 must answer:

Question 1

Can the model reveal the networks between the members in the five organisations that comprise the CEDA? This question is important because it resolves a wider and current uncertainty over the origins of the discipline of anthropology in Britain and the extent of Quaker involvement.

⁴⁹⁴ If such a database and related technologies could be built for the CEDA and made universally available, then it would be possible to extend that database by using it to also record person data on others more widely of interest to those studying the history of colonialism, and especially where information was not (and might never be) otherwise digitally accessible.

Question 2

Can the model examine Quaker-to-Quaker relationships and how these relationships supported the Quaker members of the CEDA during the forty years of its life?

Question 3

Can the model reveal the key networking role played by the Quaker Thomas Hodgkin MD (1798–1866) from the beginnings of the CEDA in 1830 up to his death in 1866?

The argument of this thesis is twofold, the first addresses a concern about the institutional and infrastructural support afforded to EBP (the subject of the preceding chapters), the second is Project Seven as a demonstration of the use of EBP in practice (this chapter). The demonstration here comprises an examination of a social network 1830-1870 and identifies the roles that Quakers played within that network. The claim is made here that Quakers played a foundational role in the early development of the institutionalisation of anthropology in Britain. That claim can (and is) shown by the P7 data analysis to be reasonably made, but the claim can only be substantiated by detailed supporting evidence found in manuscripts in Quaker archives. This supporting archival work took place, and the claim made by the data is supported by the manuscript research. This research appears in Appendix 3 and is a necessary support to the P7 Project report in Appendix 7.

6.2 Project Seven background

This chapter describes and analyses the Project Seven Report; see Appendix 7. P7 was an EBP DH project undertaken by the author of this thesis. I worked unfunded and without institutional support as a Independent Researcher, but with technical support from a

University of Birmingham Research Software Engineer (the RSE) and the assistance of the Secretary of the Quaker Family History Society (the Genealogist), who identified Quakers among the HDDT database records and found the family networks for them. The project ran from May 2019 until its completion in January 2024. It was devised by me after meetings with colonial history expert Professor Zoë Laidlaw in Oxford and London in 2018. These meetings discussed my research into the CEDA 1830–1870 and how that initial research related to Laidlaw’s own ongoing research for a forthcoming publication on the life of Thomas Hodgkin MD (Laidlaw 2021). The discussions led to a challenge: Could I improve digital support for colonial historians? We noted that colonial historians habitually research among the same sources and in the same UK archives,⁴⁹⁵ as indeed both I and Professor Laidlaw recently had.⁴⁹⁶

The background to the conversations with Professor Laidlaw is that I began a PhD research project in 2016 to better understand Quakers’ roles in supporting the CEDA from 1830 to 1870. A literature review led to locating suitable primary sources for the research, during which the archivist at the Royal Anthropological Institute (RAI) introduced me to the RAI’s in-house database (metadata) of members of the Ethnological Society of London (ESL), Anthropological Society of London (ASL) and Anthropological Institute (AI). In exchange for an export from the database of the members of the three societies from 1830 to 1870 in the form of an Excel spreadsheet, I agreed to read eight reels of microfilm of the documents of

⁴⁹⁵ ‘In addition to rendering the network of interconnections between the people obscure and disconnected, traditional practice is inefficient and costly, as professionals work in isolation from one another to establish the identities of the same persons, families, and corporate bodies, often duplicating this labor-intensive work. Walt Whitman’s papers, for example, are distributed among more than seventy repositories. In such cases, repeating biographical or historical information in two or more finding aids is an unnecessary and avoidable duplication of effort; a single creator description could be shared among all relevant finding aids (Pitti et al. 2015, 79-80).’

⁴⁹⁶ These are: The Bodleian in Oxford, The Royal Anthropological Institute in London, the Wellcome Institute in London, the National Archives, the British Library.

the APS held at the RAI, and to extract from them members' and financial supporters' names and their respective dates. The relevant data from the microfilm was hand copied onto an Excel spreadsheet. The APS data collected was later incorporated into P7 and a copy given to the RAI, who wished to incorporate the APS memberships within its founding societies database.

Several visits to Quaker archives in London revealed the names of the members of the Quaker Committee on the Aborigines (QCA), a society which pre-figured the APS and led directly to its formation. All of the names and their dates, collected from the three sources, combined with the memberships of two later societies, together comprised the CEDA memberships (QCA, APS, ESL, ASL and AI). Extensive research among the manuscripts at Friends House London included detailed analysis of the activities of the QCA and that appears at Appendix 3.

The experience of data acquisition led to me agreeing with Professor Laidlaw to determine a way of digitally sharing data gathered by colonial historians themselves, independent of any archive involvement. I agreed to at least make my own research and model (HDDT) open source and accessible to others as a template other researchers could consider using.⁴⁹⁷ I agreed in a wider context to consider the extent to which the digitisation of EBP (if it was done methodologically and efficiently) could systematise and structure the data underpinning researches into PHL.

In May 2019 a University of Birmingham RSE provided me with training and project technical advice in an initial agreement that ran from May to October 2019. This embraced data cleaning, data matching and data management, and the associated design and build of the

⁴⁹⁷ Researchers do not habitually share their sources.

expected database. Support included hands-on bi-weekly training sessions on how to use Visual Studio Code, SQLite, GitHub and DBeaver in a data model context, as well as Visual Studio Code for code scripting of the building and populating of the database. Additional independent online training was undertaken, especially with LinkedIn, the Institute for Historical Research (IHR), Data Camp, StackOverflow and YouTube. From April to October 2019, steering for the project was kindly provided by Jonathan Blaney and Martin Steer of the IHR, (who later, in 2021, along with Jane Winters and Sarah Milligan authored, *Doing digital history: a beginner's guide to working with text as data*). Two Steering Group meetings were held, one at the IHR in London and one at the University of Birmingham. By December 2019 I had decided that the project should use SQLite to build and populate the database itself, Gephi for network visualisation because of its granularity and its dynamic data visualisation capability, and Python JNB for data analysis and reporting because of its universal adoption and its ability to handle code, database analysis, text and images. It was found that JNB interfaces well with both SQL and Gephi. A further phase two agreement with the RSE then followed, which ran from December 2019 to July 2020. It provided monthly training and technical advice sessions on integrating JNB and Gephi within the phase one database and data management technologies suite. The complete project suite of Excel and open-source technologies comprised:

- Excel (CSV)
- Visual Studio Code (VSC)
- GitHub
- SQLite database (SQL)
- DBeaver (DB)

- Jupyter Notebooks (JNB)
- Gephi

The combined suite of technologies to be used were then called the Human Data Digital Toolkit (HDDT) and the project was named Project Seven to reflect the seven stages of the digital research process that arose from planning the practical work required for the project:

1. Finding and identifying (by location) Evidence Based Prosopographical Information held in a variety of primary sources, in a variety of forms and at several locations (EBPI).
2. Making representative digital records of collected EBPI (EBPD).
3. Cleaning and matching the collected digital records and combining them into one manageable dataset (Excel, VSC and GitHub).
4. Making and populating a database (HDDT).
5. Analysing data both visually (Gephi) and statistically (JNB).
6. Reporting data and analytics findings (JNB).
7. Saving the data, the project and the findings for others to use in the University of Birmingham Institutional Research Archive (UBIRA).

The project was first presented as a concept in October 2019 at the IHR. In May 2020 the Genealogist undertook (over several weeks) to (1) identify Quakers in the database records and (2) collect data on the family relationships between them. The approximately 600 Quakers and their family relationship data he found were then added to the project database, and their family relationships added as attributes. In June 2023 P7, with Quaker data incorporated, was presented to the Quaker Studies Research Association.

The balance between focusing this thesis on the model and/or the philosophies behind it required careful thought. Post-project completion review discussions with Professor Charles

van den Heuvel of the Huygens Institute Amsterdam took place in September 2023, which led to me placing greater emphasis on Information Science and the the critical examination of EBP. The discussions also included the merits of the concept of a National Authority Index (NAI) and the material contribution that it could make as a part of an EBP system to improve research affordance into PHL.

Following the discussions with Professor van den Heuvel, and to further assess the proposed NAI-UID concept, the project outline, the importance of EBP in the study of PHL and the utility of a NAI-UID in structuring and ordering EBP were discussed in October 2023 with Mark Bell, Senior Digital Researcher at The National Archives (TNA). In January 2024 outline discussions also took place with Denise Colbert, CEO of the General Records Office (GRO). The P7 project was then finalised and presented in January 2024 at the IHR.

6.3 Thomas Hodgkin MD (1789–1866)

The leader of the QLG was Thomas Hodgkin MD,⁴⁹⁸ born in 1798 in Pentonville, London, to two Quakers, both with families historically connected to the early development of Quakerism: John Hodgkin, a teacher, and Elizabeth Rickman, sometime governess to the Gurney children from Earlham, Norfolk. Thomas was the third child and the first to survive. His birth was quickly followed by that of his brother John Junior in 1800.⁴⁹⁹ Thomas was

⁴⁹⁸ The literature widely acknowledges the importance of Thomas Hodgkin MD not only in the medical field. He had a life-long commitment to relieving the plight of aborigines, he was one of the founders of the APS and a committee member of the ESL, from the formation of these societies until his death in 1866. See (Laidlaw 2007), (Laidlaw 2021), (Edmonds Penelope and Laidlaw Zoë 2019), (Lester and Dussart 2014), (Twomey 2018), (Sera-Shriar 2013), (Stocking 1987) and (Rosenfeld 2000).

⁴⁹⁹ This section of the thesis is a brief summary of the biographical accounts of Hodgkin taken from (Kass 1988), Laidlaw (Laidlaw 2021) and the *Dictionary of National Biography*,

home educated and early developed a scientific bent, when he constructed successful home experiments in electricity. At age 18 he was apprenticed to William Allen, apothecary at Plough Court. Allen was a member of several newly formed societies: the Linnean, the Royal Institution, the Geological Society and Guy's Hospital Physical Society, where he lectured regularly.

In 1819, Hodgkin authored an unpublished 'Essay on the promotion of civilisation',⁵⁰⁰ expressing his concern for the plight of North American aborigines.⁵⁰¹ His writing was informed by his meetings with Allen's 'Indian' friend John Norton and his own family's activities assisting other North American 'Indian peoples'.⁵⁰² Through this exposure, he developed an interest in the science of ethnology and the search for a scientific explanation of racial and cultural differences. As a result, he was soon to become a firm friend of the Quaker James Cowles Pritchard, widely recognised as the father of ethnology in Britain.⁵⁰³

<https://www.oxforddnb.com/display/10.1093/ref:odnb/9780198614128.001.0001/odnb-9780198614128-e-13429?rskey=lqirBh&result=3> (Accessed 28 October 2024).

⁵⁰⁰ <https://wellcomecollection.org/works/whqj34za> (Accessed 28 October 2024).

⁵⁰¹ 'A comparison between ancient and modern times, as far as relates to the influence which civilised nations have had upon the uncivilised [shows] that in the last 500 years, those under the name of Christians, have done far more to degrade, corrupt and exterminate their uncivilised fellow creatures than all the heathen world, since the creation of man. Wherever they have gone they have introduced new vices and new diseases.' Extract from the 'Essay on the promotion of civilisation', quoted in (Kass 1988, 39).

⁵⁰² 'Hodgkin's particular concern for the Indians sprang from two sources. First was the traditional benevolence of the Friends and other sympathetic observers in England and America. Hodgkin himself had met Indians who had suffered from white men's deeds. William Allen's Mohawk friend, John Norton, had received a generous reception from the Hodgkin family while they still lived in Pentonville. Other groups of Indians, hoodwinked into coming to England on false pretences only to be exploited by charlatan showmen, had been assisted by the Hodgkins and their circle in Tottenham. As Hodgkin wrote to John Norton, "from the time that I was first capable of reflecting on the subject few circumstances have excited my interest in so lively a manner, as the signal injustice and cruelty which the natives of your continent have too generally received from Europeans and their descendants. I was grieved to understand that even that intercourse of friendly character frequently had a tendency to corrupt and degrade the character and accelerate the extermination of your noble race. I longed to exert myself in a countervailing direction and was anxious to devote my abilities to your cause"' (Kass 1988, 39-40).

⁵⁰³ 'In addition to humanitarianism, Thomas had a second, equally important, reason to devote himself to the preservation of the Indians. He had begun to share the view of late 18th and early 19th century anatomists, ethnologists, and natural historians that racial and cultural differences required a scientific explanation. Annihilation of the uncivilised races would be an "incalculable and irretrievable loss" to those who wish to

Thomas's medical training at Plough Court (Kass 1988, 18-19) and work assisting Allen with his lectures at Guy's Hospital led to a distinguished medical career at Guy's (Kass 1988, 53-65) and later at St Thomas' Hospital (Kass 1988, 329-337). He also played a prominent part in the formation of the University of London (Kass 1988, 265-266). His medical career is notable for his work on the disease of the lymph nodes which bears his name (Kass 1988, 207-220), and he became commonly known as Dr Hodgkin even outside of the medical world.

While gathering data for this research project, I discovered that Hodgkin founded the APS at Ratcliffe Quaker Meeting House on 8 June 1837 (see Appendix 1) and he later became one of the founders of the ESL. He also had leadership roles in the Royal Geographical Society, the British Association for the Advancement of Science and the early British Medical Association. All of these roles presented opportunities for an unusually high level of activism and social networking.

6.4 Quaker roles in the Quaker Led Group

The QLG was active as a network from 1830 until 1870 and the extent of the group was around 3000 persons, considered over the entire period, including 600 Quakers. The group members worked consistently and collaboratively together, engaging in a set of common

study the history of early man as well as the traditions "still preserved amongst some uncivilised nations". These ideas were leading to the science of anthropology, both as a study of the physical characteristics of the varieties of man and as an examination of cultural differences. As an observant Quaker, Thomas was not prepared to consider the possibility that the origin of man might differ from the biblical explanation but he was already enough of a social scientist to appreciate the value of accumulating data on these questions. He keenly recognised the need to "save from oblivion numerous facts and traditions at present only preserved in the memory of some of those perishing nations, which, when once lost the most persevering research will never be able to recover" (Kass 1988, 40).

purposes, and the Quaker members had a leading and sustained role over the entire period in maintaining the wider group's viability. P7 shows (Appendix 7.5.18) that the Quakers arguably drove the QLG's activities through the group's support for the plight of aborigines and its later work in institution building in anthropology.

Quakers in Britain in the first half of the nineteenth century used their faith-based networks as a support and also as an engine of change when working in political arenas, especially when working closely with non-Quaker friends,⁵⁰⁴ and these non-Quaker friends in turn brought with them their own complex social networks.⁵⁰⁵ Working in complex interconnected webs of social interaction, individual members were at times strongly or weakly linked, but P7 shows that the social networks they built and maintained nevertheless were dynamic and effective. Political networks of this type were the dominant and universal mode of social engagement among elites throughout the nineteenth century, and they provided the prime arenas for social action among elites.⁵⁰⁶

Thomas Hodgkin MD appears in the manuscript sources as one of the founding members of the QCA, a Quaker Committee of Enquiry established in 1831.⁵⁰⁷ The QCA, while firmly

⁵⁰⁴ The use of friends (uncapitalised) is not to be confused with Friends (capitalised) which is the term with which Quakers customarily refer to each other.

⁵⁰⁵ 'The strong family and communal links that existed within both Quakerism and Anglo-Jewry, which are apparent in their business transactions, also operated in the election procedures in the Royal Society' (Cantor 2005, 110).

⁵⁰⁶ See (Lubenow 2015).

⁵⁰⁷ The QCA was a Quaker Committee of Enquiry. A committee formed by, and exclusively manned by, Quakers, it met, performed its enquiries and reported its findings and recommendations to the Quaker Meeting for Sufferings, the standing committee of London Yearly Meeting which was the National Assembly of Quakers in Britain at the time and the highest level of national assembly (which the QCA frequently addressed directly). The QCA's remit, rules of engagement and characteristics were those of Quakers in Britain. The committee was formed to explore and take up a 'concern' among Quakers, initially to consider promoting the Gospel among the aborigines (prompted by similar actions popular at the time among other evangelical churches). But it soon changed its remit to instead take up a philanthropic concern deriving from the group's increasing awareness of the plight of aborigines. Therefore, what began as a Quaker Committee of Enquiry to consider promoting the Gospel to aborigines quickly transformed into the Quaker Committee on the Aborigines, concerned about the plight of the aborigines and their relief. (See Appendix 3.3.)

evangelical in its founding concern, soon afterwards extended that concern to take into account a new philanthropic mission,⁵⁰⁸ perhaps suggested by the committee's awareness of the plight of North American aborigines in North Carolina (see Appendix 3.8). The learning from that exercise (and other accounts arriving from other Quaker colonialists) would immediately and significantly influence the reforms recommended in the parliamentary Select Committee on Aboriginal Tribes (see Appendix 3.13), led by Thomas Fowell Buxton supported by his extensive Quaker family (Laidlaw 2004). The QCA published the *Report of the 1834–37 Select Committee on Aboriginal Tribes* (Library of the Society of Friends. LYM Meeting for Sufferings 1837) in which at least two of the QCA members actively participated.⁵⁰⁹ The QCA's support for and driving of the work of the Select Committee led to the establishment of the APS at Ratcliffe Quaker Meeting House in 1837 (see Appendix 1), which at its inception was heavily influenced and staffed by QLG members. The APS then took up the issues raised by its founding committee and lobbied vigorously in the cause of the plight of the aborigines.⁵¹⁰ This concern of the APS for the plight of the aborigines endured for the remaining years of the nineteenth century.⁵¹¹

Alongside this work, P7 shows that the QLG also worked to bring about the establishment of the ESL in 1843 (see Chapter 6.10), thus beginning the work of institution building in anthropology in Britain. In all these activities the QLG used both the group's colonial

⁵⁰⁸ 'The deplorable condition of the heathen and the degrading circumstances under which they are living have been felt at this time, as well as in former years to be truly affecting. And although no way appears to open for the Society to adopt any specific measure in order to communicate to them the knowledge of the truths of the Gospel, we earnestly recommend their benighted condition to the Christian sympathy and frequent remembrance of all our members' (Library of the Society of Friends 1833, 6th Month 5, 394-398).

⁵⁰⁹ See (Laidlaw 2004) for a discussion of the roles of Anna and Priscilla Gurney and the wider Gurney family and other Quaker friends in researching and drafting the Select Committee report.

⁵¹⁰ Thomas Hodgkin's handwritten notes of the inaugural meeting of the Aborigines Protection Society were a considerable find (Hodgkin 1837). (See Appendix 1.)

⁵¹¹ The APS would merge with and be subsumed by Anti-Slavery International in 1919.

intelligence together with its religious and philanthropic leadings and leanings (Laidlaw 2021), which were at that time indivisibly wedded to the QLG's emerging scientific interests in anthropology (Kass 1988). However, the journey from evangelicalism to colonialism to reform to philanthropic science was not unidirectional: the flow went both ways.

Membership of the QLG varied from year to year, from as few as 50 and sometimes as many as 250 individuals, but always with a large Quaker component. P7 charts the QLG from its formation in 1831 to its end in 1869, when the QLG's ESL merged with the ASL to form the AI, which soon became the RAI, in 1871.

The QLG included among its members:

- The Pease, Sturge, Hoare, Barclay, Cunningham and Fry families and their considerable networks of Quaker service.
- Thomas Fowell Buxton and his related Quaker family (the Gurneys) and their political networks.
- James Cowles Pritchard and his Bristol/Edinburgh Quaker medical network.
- Thomas Hodgkin and his Quaker medical and scientific networks, through his friendship and collaboration with Pritchard, his faith-based network with the Gurneys, Peases and Sturges and his scientific networks relating to his medical practice.
- Richard King and his shared Quaker networks (with Thomas Hodgkin).

Regrettably the QLG had limited success in its endeavours to alleviate the plight of aborigines. From a Quaker perspective, this was perhaps because the Quaker members

found that philanthropic actions alone could not satisfy the challenge of their evangelical leadings. They also struggled to extend their influence throughout the British colonies to an extent necessary to relieve the plight of aborigines. The challenges the QLG took up inevitably stretched needs and resources too far across the disparate British colonies ever to allow the society's aims to be realistically achievable. Nevertheless, the QLG played a significant (and at times a leading) role as the vocal conscience of the British nation about the treatment and recognition of aborigines throughout the British colonies from the 1830s onwards. It achieved this through its involvement with the APS,⁵¹² and Laidlaw suggests that the APS may well have exerted political influence in ways too subtle for the official records of the time to fully recognise.⁵¹³

It is, perhaps, the apparently limited success of the QLG in its initial mission to relieve the plight of aborigines throughout the British colonies, contrasted to its organisational success at home in nurturing the emergent discipline of anthropology, that makes this group particularly worthy of academic attention and scrutiny. Despite its failure to achieve its lofty philanthropic and evangelical objectives, the QLG itself nonetheless exhibited significant cohesion, durability and adaptability throughout its long life and, if there were no great successes, there were indeed many small ones. It performed consistent and vigorous lobbying activity at the highest levels of British political life starting in the 1830s, and it maintained that political pressure well into the 1870s through its publishing arm and its

⁵¹² 'Not all individuals within a network carried the same weight: not only in the sense of asymmetric relationships but also in terms of the number and strength of their connections. Thomas Fowell Buxton, for example, the parliamentary anti-slavery campaigner and co-founder of the Aborigines Protection Society, had many contacts in British colonies, among missionaries, church leaders, scientists, and administrators, as well as a broad range of connections – through his family, religion, and parliamentary status – in Britain. Buxton therefore was not only a central node in the network of humanitarians, but was also able to integrate this network within several others' (Laidlaw 2005a, 15).

⁵¹³ 'Henry Hall, for example, started his account of the Colonial Office in 1836, because the absence of minutes on documents before then made any "detailed" study of an earlier period "useless"' (Laidlaw 2005a, 49).

regular publication the *Colonial Intelligencer* (see Appendix 3.13).⁵¹⁴ Therefore, although the group was not borne along by its successes, it nevertheless continued, often with doggedness and driven by the passions of the group's members, for the over-riding mission they shared, to lobby for the rights of aborigines.

In this light, the QLG's contribution to institution building in anthropology might perhaps be seen as a surprising consolation prize, or simply an opportunity that emerged alongside the QLG's intentionally greater philanthropic endeavours, and is best understood within the age of disciplinisation sweeping Britain.⁵¹⁵ But it was a significant gain nonetheless, and it is of historical interest how the QLG, weaving social webs between the four historical themes (see Section 6.6), produced a flag of one cloth, which on one side shows the colours of political action in support of aborigines and on the other side those of institution building in anthropology. An important driver for this thesis is to explore the durability of a demonstrably part-successful and part-unsuccessful political/academic group that achieved much, and did so over a lengthy period of time, almost solely by the busyness of its Quaker members' networking activities.

6.5 Issues in defining the QLG members

⁵¹⁴ <https://catalog.hathitrust.org/Record/011725668>. (Accessed 18/12/2024)

⁵¹⁵ 'A completely new understanding of encyclopedias emerged only at the beginning of the nineteenth century. The quality of encyclopedias no longer resided in the fact that singular scientific truths obtained their place in science only by being positioned in such publications. Rather, encyclopedias became reflexive; they described themselves as the science of science — which presupposed that science existed independently of encyclopedias. The latter thereby became an institution for observing science. Simultaneously, there was an increase in the use of organic metaphors to describe specific sciences and the connections among them (see Stichweh 1991b). There is here an obvious tendency to perceive a science in a new sense, as a living organism independent of external interventions aimed at bringing about order' (Stichweh 1992, 6).

It is not easy to define the QLG members because the group was loose and it adapted to changes over time. Unlike the Quaker members, the wider QLG did not have an organisation or membership criteria, any rules of membership, or any of the other usual ways of qualifying or demonstrating membership. However, a strong sense of inter-relationship among the group emerges from the manuscript sources themselves, and this is revealed in the detailed P7 Case Studies based on the prosopographical data embedded in the manuscripts.

The many references to these persons and their connections as they appear in secondary sources can only hint at the level of networking that actually took place. This is because the relevant secondary literature is both numerous and diverse. QLG members were also frequently members of several other networks, therefore tracing the QLG through secondary sources alone requires careful attention to separate the QLG from discussions centred on other networks. The older secondary literature usually only considers a few key individuals from the QLG at a time (and mostly the famous men among them),⁵¹⁶ but some of the recent literature in this area adopts a broader canvas.⁵¹⁷

The collection of individuals which this study refers to as the QLG could be considered as a friendship group, but the friendships that emerge from a study of the manuscripts and recent secondary literature go well beyond a loose or general definition of friendship. The literature indicates that these friends supported each other materially and emotionally, they engaged in significant shared public actions and demonstrated a richness of cohesion, at work, at Quaker Meeting and at home – and it is this richness that binds them together. QLG members moved freely and fluidly between several learned societies and networked with

⁵¹⁶ See, for example, (Stocking 1987, 1971; Rainger 1976).

⁵¹⁷ See, for example, (Laidlaw 2004).

other religious communities in London and throughout Britain. As well as pursuing objectives derived from within their Quaker faith community and its philanthropic and evangelical concerns, the Quaker members of the group mixed their faith-based concerns seamlessly with other, more worldly concerns and objectives, and progressively included interests more characteristically deriving from the QLG's shared enthusiasm for scientific interests, as (Kass 1988) show. The QLG members, Quaker or otherwise, routinely found ways to combine and exploit sometimes seemingly unrelated interests with surprising efficiency and effectiveness.

This thesis chooses a loose definition in determining the membership of the QLG, in considering who is, and is not, and when, a member.⁵¹⁸ This study defines QLG members to be primarily those whose names appear in the manuscript sources as members of, and supporters of, one or more of the CEDA in Britain. QLG members also sometimes appear as prominent members of other related science-centred organisations that were emerging at the time,⁵¹⁹ and they were also occasionally found to be active members of other faith-based communities.⁵²⁰ All of the persons considered here as the QLG cannot be neatly contained within a narrow definition of Quaker membership, and occasionally discernment is necessary in establishing group membership.⁵²¹

⁵¹⁸ For a similar consideration of issues in too closely prescribing definitions in social group identification, see (Cantor 2005, 6.3.3).

⁵¹⁹ 'At the Royal Society rooms or the annual meeting of the BAAS, Quakers and Jews rubbed shoulders with Anglican clerics, Methodists, Unitarians, Catholics, agnostics, and atheists. In so doing they not only enhanced their own respectability by participating in the burgeoning study of science, but also met and liaised with scientifically literate doctors, politicians, aristocrats, and even Anglican clergymen' (Cantor 2005, 103).

⁵²⁰ The QLG included members from Anglican, Methodist and other largely evangelical churches.

⁵²¹ 'Who is to count as a Quaker or as a Jew? – At first sight this question may seem easy to answer. However, in both cases there are a number of difficulties. Turning first to the Society of Friends, during the period covered by this study, the vast majority of members were Quakers by "birthright"; that is, their parents were Quakers. Yet a significant proportion of "birthright" Quakers deserted the sect or were disowned for any number of reasons: for example, for failing to attend meetings, for parenting an illegitimate child, or for

6.6 The Centres for the Emergence of the Discipline of Anthropology in Britain

For approximately forty years the friends who made up the QLG worked closely together and in various combinations, and their actions significantly influenced five collectives that addressed the plight of aborigines and/or were precursors to the emergence of the discipline of anthropology in Britain:

- Quaker Committee on the Aborigines, QCA (1831–1846)
- Aborigines Protection Society, APS (1837–1848)
- Ethnological Society of London, ESL (1843–1848)
- Anthropological Society of London, ASL (1861–1869)
- Anthropological Institute, AI (1871), a merger of ESL and ASL

This thesis refers to these organisations as the Centres for the Emergence of the Discipline of Anthropology (CEDA).

breaking with any of a number of Quaker tenets. Prior to the 1860s the most frequently cited reason was marriage to a non-Quaker. Elizabeth Isichei has estimated that of mid-19th-century Quakers, “between a quarter and a third of all (Quakers) who married at all” married out, and would therefore be disowned. This severe haemorrhage threatened the very existence of the society, and from 1861 a number of changes were implemented, the most important being repeal of the proscription against intermarriage. Moreover, in each generation some who were not of Quaker parentage were attracted to the movement and, through conviction, were accepted into the Society. Prior to the 1860s, these recruits generally accounted for only a small fraction of the total Quaker population, but thereafter increasingly became the norm. Some people who were not Quakers nevertheless regularly attended meetings. However, as these attenders were not subject to Quaker discipline, they have not been considered here, except in a very few specific instances’ (Cantor 2005, 358).

6.7 The Quaker Committee on the Aborigines

The QCA began its life as a Quaker Committee of Enquiry (1831–1833), a committee formed by and exclusively staffed by Quakers. The Quaker Committee of Enquiry met, performed its enquiries, and reported its findings and recommendations to the Quaker Meeting for Sufferings, the standing committee of London Yearly Meeting, which was the national assembly of Quakers in Britain at that time. The Committee of Enquiry's successor, the QCA, also frequently addressed the Quaker national body directly. Both Quaker committees worked under a remit and rules of engagement established and monitored by Quakers in Britain under the aegis of the Meeting for Sufferings. The QCA published extensively; see Appendix 3.14.

The Committee of Enquiry was formed to explore and take up a 'concern'⁵²² among Quakers, initially to consider promoting the Gospel among the aborigines, prompted by the popular missions of other dissenter faiths at the time, usually more enthusiastically evangelical churches. The Committee of Enquiry was short-lived and does not appear to have been successful. Two years later in 1833, it changed its remit to instead take up a philanthropic concern that derived from the group's increasing awareness of the plight of aborigines which it had gained during the Committee of Enquiry's term of office. Therefore, what began as a Quaker Committee of Enquiry to consider promoting the Gospel to aborigines quickly transformed into the Quaker Committee on the Aborigines, which was

⁵²² A Quaker concern: 'An issue or idea that Quakers feel spiritually led to take action on: the leading could be for an individual to act, or for Quakers to take action collectively. Quakers who feel they have an idea that might be a Concern will ask their meeting to help them test this.' <http://centralenglandquakers.org.uk/quaker-terms/concern/> (Accessed 18/10/20324)

concerned about the plight of the aborigines throughout the British Overseas Territories and their relief.

6.8 The Select Committee on the Aborigines

The House of Commons Select Committee 1834–1837⁵²³ was called to investigate the condition of aboriginal tribes in the British Settlements, reporting its findings and recommendations to Parliament. The Committee's remit, rules of engagement and characteristics were typical of those of the many other Parliamentary Select Committees of the 1832 reform Parliament. The work of the parliamentary committee was directly and substantially supported by Quakers (especially the Gurney family in Norfolk) and the Quakers involved were themselves connected to the Quaker Committee of Enquiry and the QCA (Laidlaw 2004). The report of the parliamentary committee was afterwards published and taken up by the APS.

6.9 The Aborigines Protection Society

The Aborigines Protection Society was a Quaker-led secular pressure group that, from its formation in 1837 until its absorption into the Anti-Slavery Society in 1909, relentlessly lobbied the Colonial Office and Parliament for the relief of the plight of aborigines throughout the British Settlements. It had a mixed Quaker and non-Quaker executive, membership and subscription list and many of its first members were drawn from the QCA.

⁵²³ <https://apo.org.au/node/61306> (Accessed 18/10/2024)

Quakers dominated the agenda, publishing and lobbying activities of the APS for at least the first thirty years of its life. The APS met monthly in London (usually at Exeter House) and reported its findings and the executive's recommendations to its members according to its own constitution. The APS's remit, rules of engagement and characteristics were similar to those of the many other secular lobbying and public opinion forming societies of the time.

6.10 The Ethnological Society of London

The Ethnological Society of London was the first intentionally academic society devoted to the discipline of anthropology. Secular by intent, if not always entirely so in its early years, it sought to be a place where those with a scientific interest in the field of ethnology could commune, share ideas and knowledge, and produce academic reports and hold academic meetings. It met, performed its enquiries and reported its findings and recommendations to the society's members according to its own constitution (it usually met monthly). The ESL's remit, rules of engagement and characteristics were similar to those of the many other scientific societies emerging at the time, its constitution being purposely compliant with British Association for the Advancement of Science (BAAS) requirements, because one of its primary aims was the inclusion of ethnology and anthropology as a BAAS accredited science.⁵²⁴

⁵²⁴ See Hodgkin's report to the 1841 Annual Report of the BAAS setting out his intention that ethnology be included as an accredited science (which prefigures the establishment of the ESL): 'Dr. Hodgkin concluded by urging, as practical means for advancing the cause of Ethnological investigation, first, the bringing home, for the purpose of being studied themselves, as well as of being made the subjects of suitable education, well-selected aboriginal youths, and especially such as have had an opportunity of acquiring knowledge, and exhibiting ability in missionary or other native schools. This plan, which need not equal in expense what is often done for other objects of zoology and for botany, might be facilitated by the union of individual

A key presence and leading activist on all three of the above groups was the Quaker Thomas Hodgkin MD, a lifelong friend to James Cowles Pritchard, a leading academic who is now widely held to be the ‘father’ of anthropology in Britain.⁵²⁵ Pritchard was born into a Quaker family but resigned Quaker membership in order to take up a scholarship at Oxford, although he remained closely associated with Quakers throughout his life, especially Hodgkin. Pritchard was to become the first chair of the ESL and died soon after the society’s formation. Hodgkin shared Pritchard’s interest in ethnology, himself lecturing and publishing on ethnologically related matters from the early 1820s. Hodgkin used the established QLG network from the two earlier Quaker committees and the APS, both to bring Pritchard from Bristol to London and then to lead in the formation of the ESL. The relationship between these organisations and their importance in understanding the emergence of the discipline of anthropology in Britain was first noticed and taken up by Ronald (Rainger 1976).

6.11 The QLG and the emergence of the discipline of anthropology

In addition to exploring relationships between Quakers and the QLG in their seeking to relieve the plight of aborigines, this thesis focuses on the QLG’s pursuit of the discipline of anthropology because, although this group’s primary role – its concern for the plight of aborigines – has already been examined in the secondary literature (see (Laidlaw 2007, 2005a, 2001, 2004), the journey from Quaker concern to the setting up of the ESL has not.

contributors. Secondly, rendering personal and pecuniary aid to the Aborigines’ Protection Society, the objects of which were neither of a party nor of a sectarian character, but were solely directed to the preservation, amelioration, and study of the feeble races of mankind, amongst which those related to British colonies occupied the chief place, Dr. Hodgkin observed, that the objects pursued by this Society furnished subject matter not merely for the Zoological, but also for the Medical and the Statistical Sections.’

<https://www.biodiversitylibrary.org/item/104191#page/33/mode/1up>, p. 52 (Accessed 10 January 2025).

⁵²⁵ See the section on ‘The Remarkable Dr Hodgkin’ in (Stocking 1987).

The QLG's networking among the scientific clubs clustered around the Royal Society and the BAAS, and other clubs associated with the emergence of the discipline of anthropology, has also been explored in the literature. The P7 study analyses in detail the person-to-person relationships between the members of the QLG and how those relationships, reinforced by their shared concern for the plight of aborigines, influenced and shaped the affairs of societies that brought about the emergence of the discipline of anthropology in Britain.

That the QLG intended to foster the emergence of the discipline of anthropology in Britain is not clear. An examination of the manuscript sources shows that the group instead came together in Quaker 'concern' for the promulgation of the Gospel among the 'aborigines' in the late 1820s. The QLG then spent the next decade or so trying to relieve the plight of aborigines throughout the British Overseas Territories. Learning from the experience of these efforts, some members of the QLG had emergent scientific objectives in mind, especially Hodgkin, and they used the platform of their philanthropic interest to pursue institution building in anthropology.⁵²⁶ By 1870, the QLG had quietly faded away as the future of the discipline of anthropology in Britain had by then passed from activists to academics. The long journey was complete when the ESL, having surviving the death of its first (Quaker-born) leader James Cowles Pritchard in 1848, and then husbanded by Thomas Hodgkin MD until his death in 1866, was finally absorbed into the AI, which became the Royal Anthropological Institute in 1871 (Stocking 1971). After this time, the residual

⁵²⁶ Hodgkin's paper 'The progress of ethnology' for the first *Journal of the Ethnological Society of London* (1848–1856) links his interest in ethnology with his interest in the plight of aborigines: 'A great number of curious problems in physiology, illustrative of the history of species, and the laws of their propagation, remain as yet imperfectly solved. The psychology of these races has been but little studied in an enlightened manner, and yet this is wanting, in order to complete the history of human nature, and the philosophy of the human mind! How can this be obtained, when so many tribes shall have become extinct, and their thoughts shall have perished with them!' (Hodgkin 1848). See also (Kass 1988, 183).

philanthropic work of the QLG was subsumed into other Quaker committees,⁵²⁷ leaving the APS to continue with its much reduced lobbying work to continue to try to alleviate the plight of aborigines, until it in turn became a part of Anti-Slavery International in 1919.

6.12 Quaker families

The Genealogist who assisted P7 by identifying Quakers and their relationships used commercial genealogical software and identified about 600 Quakers in the expanded P7 dataset of 3000 persons. He then used the same software to find the familial relationships between the 600 Quakers. The data gathered was exhaustive as to known CEDA memberships, but by no means exhaustive as to person attributes. Further associative connections between the members were hinted at in the data but were not collected. Neither was the dataset complete.⁵²⁸ But the enhanced dataset was judged sufficient to produce some useful learning and sufficient to offer a critique of practice in the field. (To complete an exhaustive data gathering exercise would be a lifetime's task for several researchers.)

6.13 The HDDT design challenges

Careful project planning and design were essential throughout the modelling work of P7 because it embraces metadata and both primary and secondary data from several sources,

⁵²⁷ On 6 August 1847 the Committee on the Aborigines merged with the Africa Committee and Thomas Hodgkin became a member of the merged groups.

⁵²⁸ Data for some years was irretrievably lost.

and it uses a range of technologies to clean, organise, manage and visually interpret the data. Care was taken in project design, the qualities of the data and the selection of technologies to understand (1) what happens to the integrity of data as it moves through the technology pipeline, and (2) how the technologies chosen influence project and data outcomes. Therefore, before building the HDDT model, particular care was taken to evaluate the flow of data inputs and outputs, to ensure that the HDDT would bridge them both efficiently, effectively, losslessly and openly. It was also anticipated that a considerable amount of auditing and data verification was needed all along the technology and data pipelines, and that selected audit tasks and interventions needed to be established at the outset to ensure that they were manageable and appropriate, both in terms of respect for the data and the considerable burden of data cleaning. Trial runs to practise methodologies and techniques took place on small samples of data before finalising model design. Technical exercises of this sort are time consuming because they must be thorough, and so time was included to allow for deep reflection on the inner workings of the model before key decision making and building took place. The extent to which digital hermeneutical considerations arose throughout the P7 project reflects the complexity of working in DH, even if the project is small and run by a Independent Researcher.⁵²⁹ Issues arose such as distancing: was the digital data too far removed and disconnected from its primary source, and does this lead to limits of confidence in the project data, its use and its Interpretation? This is important because DH affordances are still relatively new and, because user interfaces are impressive and 'modern' looking, they can easily suggest that outputs can be relied on with relatively little questioning (using the technology can seductively invite

⁵²⁹ 'The need to reflect on how computers influence the construction of scientific knowledge' (Romein et al. 2020, 309).

passivity on the part of the researcher). Did the requirements and constraints of the technological and methodological choices made distort data management and outputs? This was important because data management and analysis affordances change at each stage in the pipeline due to changes in technology. In spite of pre-project thinking, actually using the model resulted in discovering unanticipated subtleties in data transitions through the HDDT.⁵³⁰

6.14 Complex prosopographical networks

This study shows that the QLG during the years 1830–1870 was highly networked. To be included in the networked life of elites in Britain at this time one had simply to be accepted as a member of ‘society’, a loose but prescriptive term for middle-class gentlemen and occasionally their wives (see (Laidlaw 2001; Lester 2005; Cantor 2005). Society had its own ways of policing who was ‘in’ and who was ‘out’, through modes of politeness, education, breeding and even prohibition, as in the case of Dissenters.⁵³¹ Importantly, social

⁵³⁰ ‘There is a difference, however, between historians who engage passively with historical content in digital form when they browse the web looking for literature and data, and those who are committed to a fully digital research process. While the first will eventually produce a printed monograph, the second, still a minority, will use digitised or born-digital data, often neatly arranged in a database, analyse it with digital tools, and publish the results in the form of a website or a peer-reviewed publication supported by a dataset and code. Both categories can continue to do what historians have always done, question the origin and authenticity of a historical source by determining when it was created, by whom, for which purpose and with which means. Nevertheless, in the digital age, this has to be complemented with a more technical and mathematical understanding of digital phenomena. Besides reflecting on why a particular collection of documents has been selected to be digitised and published on the web, a historian should also be able to identify the alterations and loss of context that occur when the collection is transformed from its analogue to its digital form’ (Romein et al. 2020, 309).

⁵³¹ ‘As Jack Morrell and Arnold Thackery have stressed in their history of the BAAS, early meetings were dominated by Dissenters and broad Anglicans. Clearly, Quakers conformed to the Association’s ethos. They would have found the Whig, reformist ambience as congenial; no oaths were required and no religious tests; instead, the meetings were open to all – at least all who could afford tickets (which included the majority of scientifically informed Quakers). The Quakers who attended these annual gatherings, both men and women,

networking was how these elites communicated and bonded when out in the world, through their many and fluid alliances and associations. 'Network' and 'networking' are modern terms that Hodgkin's contemporaries would not have used. So obvious and universal was networking, however, that it did not need to be examined as a unique phenomenon. Networking for the QLQ meant active memberships in the QLQ's lobbying and/or academic societies, each with clear objectives, organisation, funds and officers. Activities included the writing and presentation of research papers, producing, promoting and reading society journals and publications, attending regular events organised by the respective societies and supporting the society officers in campaigns and promotions of the society's objectives. Networking was also how individuals went about achieving their goals in their wider political, economic and religious lives, weaving and inter-weaving networking webs to great effect.

Social networking at this time therefore goes beyond and deeper than what is usually described in discourses and historical narratives by academics and others when studying social reform in Britain in the nineteenth century, because the extent of networking in practice is beyond the scope and grasp of traditional research techniques. These accounts by historians without benefit of digital technologies (some made in the very recent past) are constrained by the practical limitations of 'in-person' working in archives and an inability to analyse mass data. Almost invariably, this results in the individuals studied necessarily being sorted by the researcher into 'groups', established by the writer to better handle data rather than reflecting structures present in the data itself. But these often artificial 'groups',

mingled freely with the other genteel ticketholders. Indeed, the association was particularly attractive to Quakers because it enabled them not only to meet other Friends, but also to encounter other respectable people, from outside the restricted Quaker community, who shared their scientific interests' (Cantor 2005, 131).

made after the event by historians, are not lived 'networks' because they have too often been manufactured long after their members were deceased. All too often these groups made much later can include individuals who in life never met and may not have even known each other. Little research has been performed into the true nature of the connectedness (or not) between these grouped individuals as they can be observed, in their own lifetimes.⁵³²

6.15 P7 Topology

A data topology for the project was devised that both bounds and frames the extent and content of the social networks to be modelled and that also frames the three Case Studies. A data topology needs to represent the thematic boundaries of a given study and to facilitate the objectives of that study. In this case the objective was to reveal the prosopographical characteristics of the 3000 activists and to make visible the relationships between them. This study adopted four 'bounds' to the data topology, each of which is in itself a theme of contemporaneous social life:⁵³³

- colonialism
- reform
- philanthropy

⁵³² 'Analyses of human social networks have a long history in both the sociological and anthropological literature (Milardo 1988). However, relatively few studies have attempted to investigate complete social networks in humans (McCarty et al. 1997), primarily due to the difficulty in estimating and defining an individual's "network" from the range of interactions that exist within everyday life' (Hill and Dunbar 2003, 54).

⁵³³ Other themes might be chosen for other (even related) historical enquiries where the subject of research is not the QLG, for example 'technology', 'diplomacy', 'international law' and 'nation statehood'. The choice of bounds for historical enquiry of course risks not providing a sufficient window on the past to adequately capture the social activity studied. Also, if the bounds are not at least some of those used by other historical enquiries, this can leave the study unable to benefit from the work of peers.

- evangelicalism

The four topological bounds make a topological frame which would contain one overarching feature of civic life that was considered to be the central theme of the study: the ‘age of disciplinisation’.⁵³⁴ This topological schema can then be represented as in Figure 6.1.

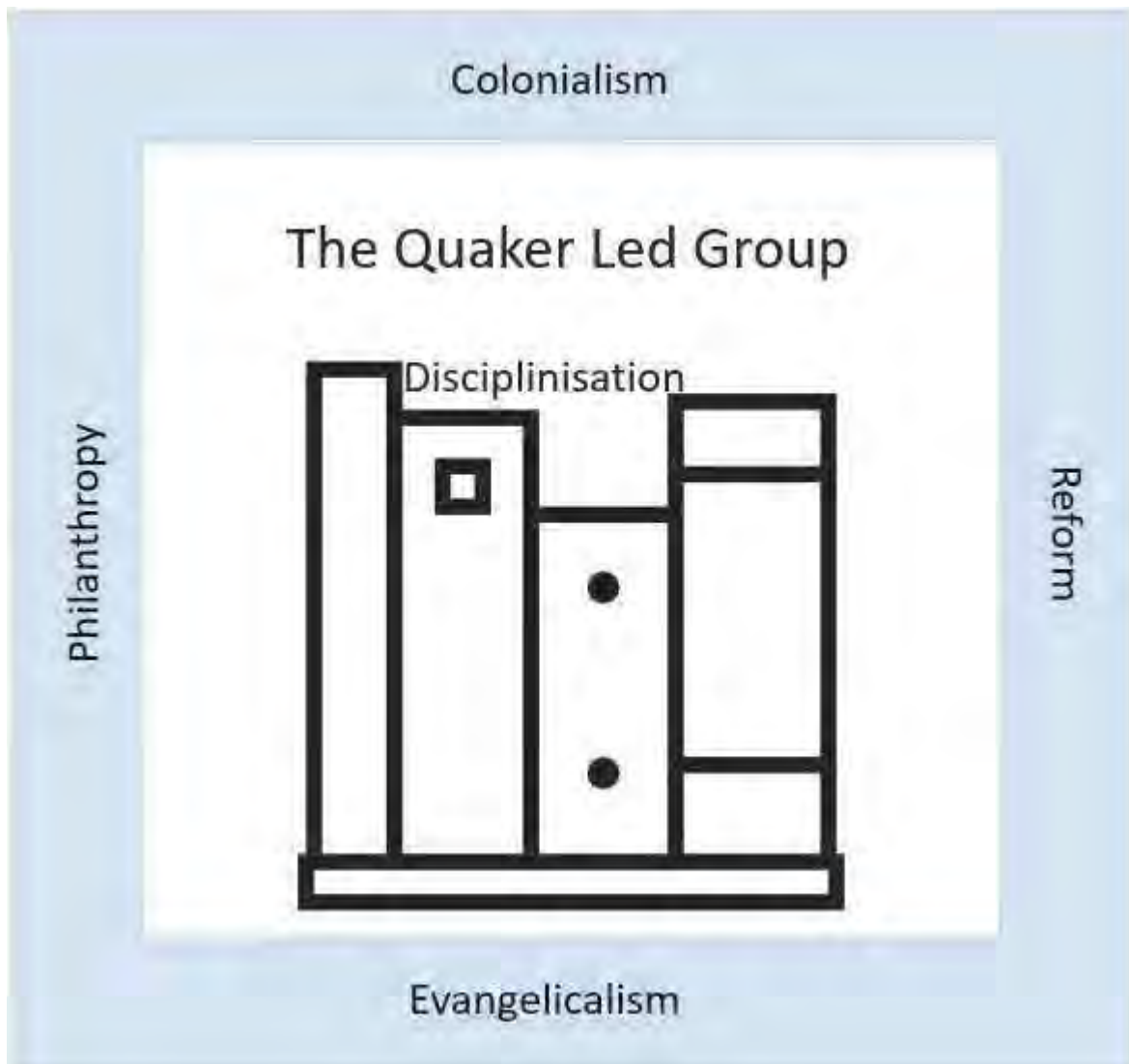


Figure 6.1 A Quaker Led Group topology

⁵³⁴ The term ‘The Age of Disciplinisation’ is used to describe the increase in record making and keeping, especially government records, which arises from about 1800. A feature of the activism of the QLG is its ability to gather data of higher quality and at a faster rate than that of the government the group sought to influence see (Stichweh 1992).

These four narratological themes could be imagined as four pillars of historical discourse that can be used to structure a study of political activism in Britain in the first half of the nineteenth century. The brickwork, then, that fills in the gaps between the pillars would be the many social/political networks, each built by overlapping and inter-connecting social groups. For the purposes of this study, only one of these social/political groups (and one not previously identified in scholarly research) is the object of this thesis.

6.16 Designing the HDDT

Four of the over-arching design objectives for the suite of technologies that would become the HDDT related to P7 execution and a final two to post-P7 reflection:

1. Locate and digitally extract data from several UK archives relating to persons in the period 1830–1870, who were concerned about the plight of aborigines and also who supported the emerging disciplinisation in the study of ethnography/anthropology in Britain before the formation of the Anthropological Institute (AI) in 1871, now the Royal Anthropological Institute (RAI). Brief prosopographical data on 3000 such persons, including over 600 Quakers, was found, sufficient data to build a project database and a digital analytical model.
2. Design and build a digital data model including a database (more properly a set of variously integrated models), to contain the data collected, to clean and organise the data, and to analyse and visualise the data, much as a colonial researcher might.

3. Integrate the data model into University of Birmingham data management systems and also non-academic open access systems (GitHub) – adopting best practice and FAIR principles throughout.⁵³⁵
4. Use the model to analyse and visualise the relationships between the 3000 persons recorded in the dataset.
5. Reflecting on the data collection exercise, research the current state of digitisation in Digital Humanities (DH) at national, archival and research project levels, to reveal the extent to which current levels of digitisation facilitate or support research into Past Human Lives (PHL) using Evidence Based Prosopography (EBP). It was also important to focus on the ‘Lone Historian’ model⁵³⁶ if this project was to be of help to often overlooked researchers working with limited funds and limited access to prestige institutions.⁵³⁷ The needs of the Lone Historian and the help available vary from those able to benefit from physical access to a well-funded and supported digital research hub and those who rely solely on digital service support.
6. Using the learning from data collecting, data modelling and visualisation alongside the investigation into current digitisation infrastructures in DH, identify structural weaknesses in research practices into PHL and recommend new directions for digitisation efforts that would lead to better research practice (especially for Lone Researchers) and that might, over time, build stronger and more user-appropriate, sustainable and reproducible data affordances across the research community, and

⁵³⁵ <https://www.go-fair.org/fair-principles/> (Accessed 4 June 2024).

⁵³⁶ ‘Existing as a separate discipline offers digital humanities scholars much needed support for pioneering interdisciplinary research and methodological debate, but at the same time weakens their ability to reach out to oft-caricatured monodisciplinary “lone scholars” who (for better or worse) continue to represent most historians’ (Blaxill 2023, 288).

⁵³⁷ Especially important in the field of colonial history, where today resources are still disproportionately found at research hubs in countries with a history of colonisation rather than colonised countries themselves.

also in the many archives and research hubs. These considerations, based on my research experience with the P7 project, would include considering the benefits of bringing together varied practitioners who share an interest in Evidence Based Prosopography Data (EBPD): family historians, genealogists, archivists and librarians, researchers in research hubs and Lone Researchers.

6.17 Building the HDDT

The P7 project assembled data from various sources and in different forms, and a group of modelling technologies were chosen by the P7 author after project scoping exercises with the RSE that took as their starting point the quantity and quality of the data, and after a desktop review of potential technologies was completed. The HDDT needed to have technical compatibility that would seamlessly integrate across the whole project data pipeline. The P7 technology requirements were tasked with both recognising the project objectives and enabling the three set questions to be addressed. The HDDT needed to:

- be able to handle the data (quantity and quality);
- allow user interventions at each stage;
- be open source;
- be popular and in common use;
- allow ease of Interoperability;
- be supported by free online learning aids;
- be supported by University of Birmingham Research Software services;
- be Independent Research Group friendly, undemanding of time and resources.

The chosen technology suite adopted by P7 comprised:

- Excel – to create CSV sheets of EBP data. Data was captured in clusters and with different qualities. At the beginning of an archival research session the frequency, volume and style of data yet to be encountered are unknown, so Excel is an appropriate data collection and crude data organising (as you go) tool, but work consolidating archival captures was needed.
- Visual Studio Code to create and populate the relational database (Figure 6.2).

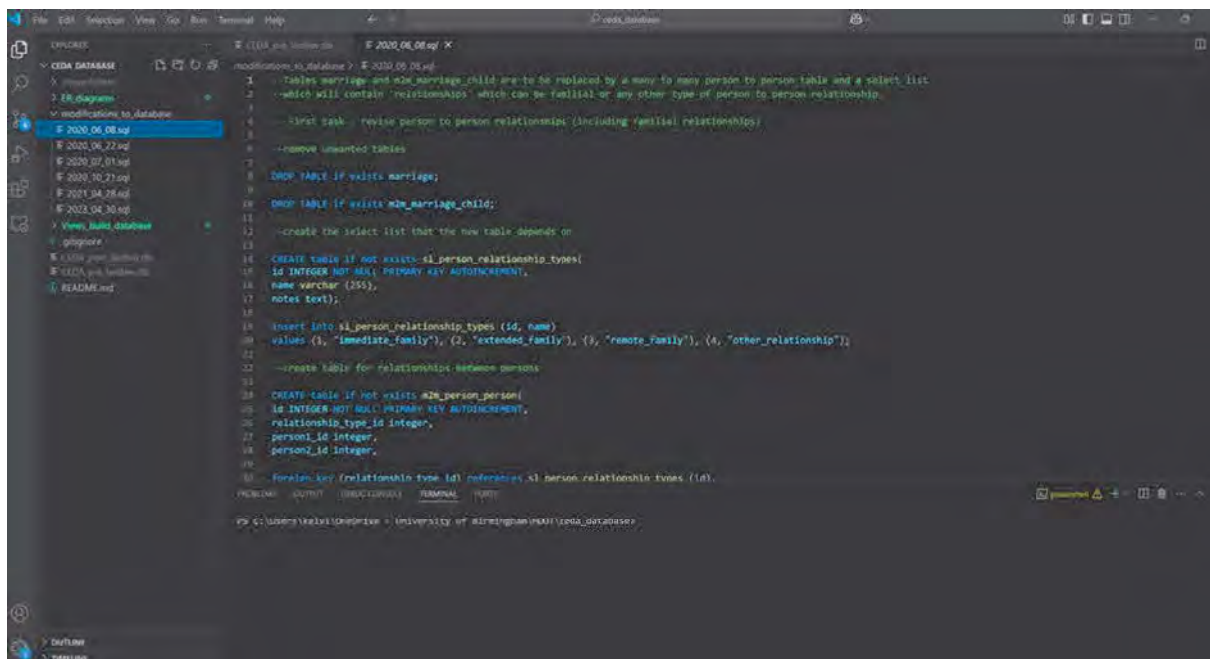


Figure 6.2 Visual Studio Code was used to build the database

- DBeaver to manage, interrogate and provide segments of the SQLite data (Figure 6.3).

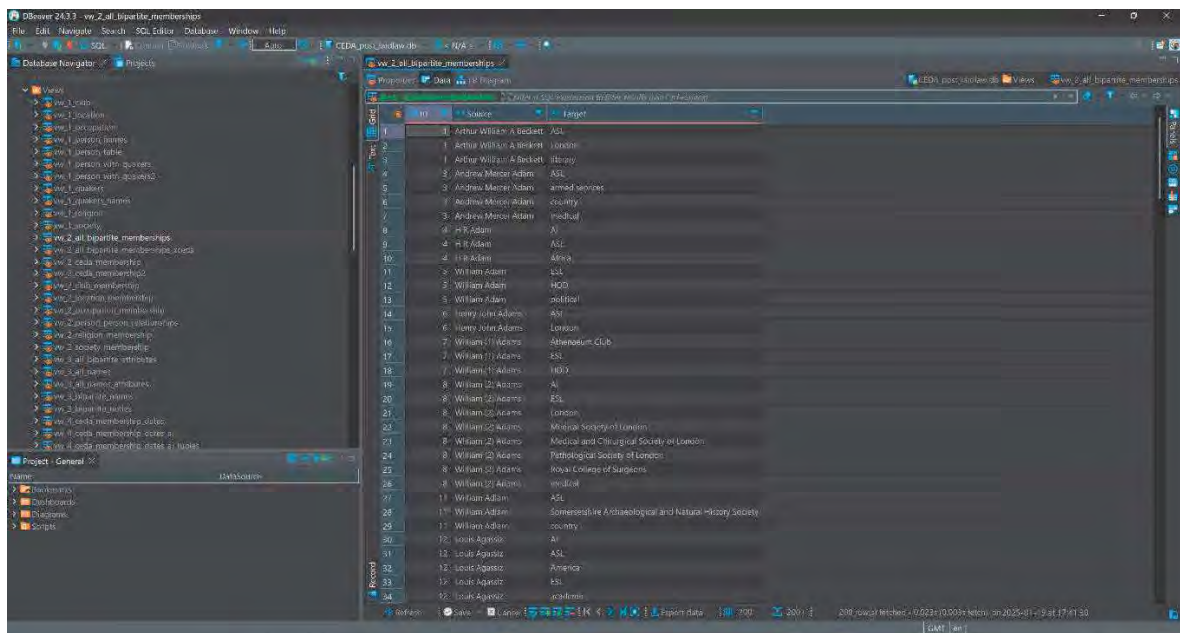


Figure 6.3 DBeaver Desktop

- GitHub for version control of both code building and data cleaning exercises. GitHub was also the repository for all of the JNB Case Study project exercises. This will be the online ‘home’ of the project code online after completion (Figure 6.4).

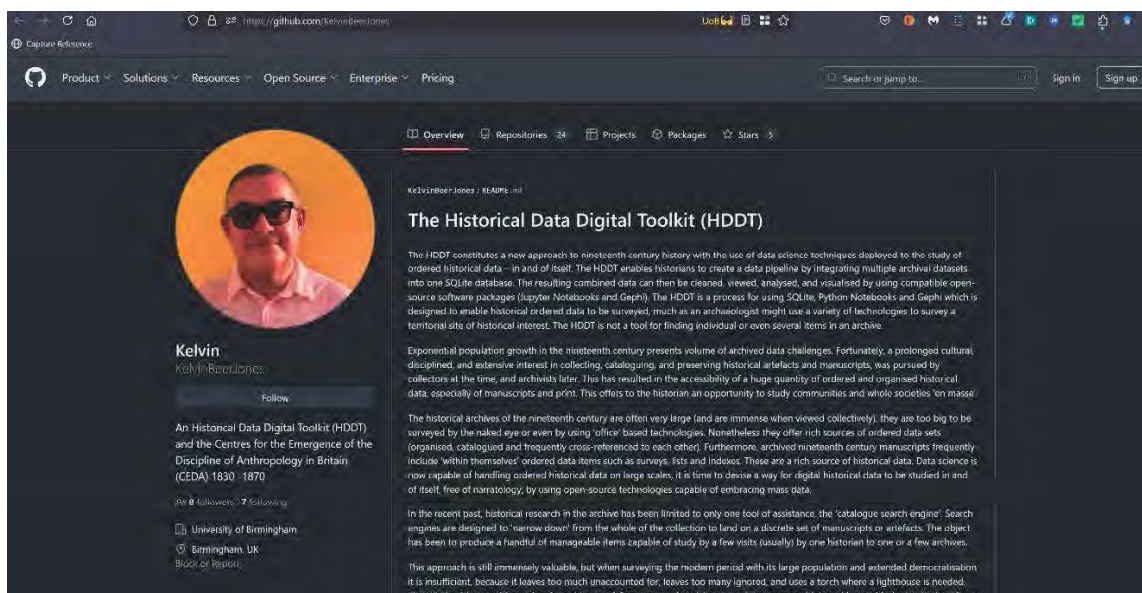


Figure 6.4 The Project Seven GitHub repository

- SQLite to house the database, allow data queries and section the data in DBeaver for export to JNB and Gephi.

- Python Notebooks for project execution using dataframes, Numpy, Pandas and Matplotlib, alongside NetworkX routines (to generate Gephi data files). JNB was able to interface efficiently with the SQLite database inputs and to export refined data to Gephi, the graph database (Figure 6.5).

The screenshot shows a Jupyter Notebook interface with the title "Bigraph nodes (Names) and edges (with attributes)". The notebook contains the following code:

```

In [33]: # Import csv
from operator import itemgetter
import networkx as nx
from networkx.algorithms import community # This part of networkx, for community detection, needs to be imported separately.
import abc

In [34]: # With open('nw_3_all_names_attributes.csv', 'r') as nodescsv: # Open the nodes csv file
nodesreader = csv.reader(nodescsv) # Read the csv
nodes = [n for n in nodesreader[1:]] # Retrieve the data (using Python list comprehension and list slicing)
# To remove the header row
node_names = [n[0] for n in nodes] # Get a list of only the node names

with open('nw_3_all_bipartite_attributes.csv', 'r') as edgescsv: # Open the file
edgesreader = csv.reader(edgescsv) # Read the csv
edge_list = list(edgesreader) # Convert to list, so can iterate below in for loop

# Create empty arrays to store edge data and edge attribute data
edges = []
edges_attributes = []

# Fill the arrays with data from csv
for e in edge_list[1:]:
    edges.append([e[0], e[1]]) # Get the first 2 columns (source, target) and add to array
    edges_attributes.append(tuple(e[2:])) # Get the 3rd and 4th columns (first_year, last_year) and add to array

edge_names = [n[0] for n in nodes] # Get a list of only the edge names

In [35]: # Print('Nodes length: ', len(node_names))
print('Nodes length: ', len(nodes))
print('Edges attributes length: ', len(edges_attributes)) # This should be the same length as edges
  
```

Figure 6.5 A Project Seven JNB

- Gephi visualisation network analysis tool for network visualisations. Network graphs display data statically (whole period 1830–1870), year by year and dynamically, time sequencing one year at a time from 1830 through to 1870 (Figure 6.6). Gephi has rich functionality with a choice of layouts, network topologies, community detection algorithms, and both node and edge views.

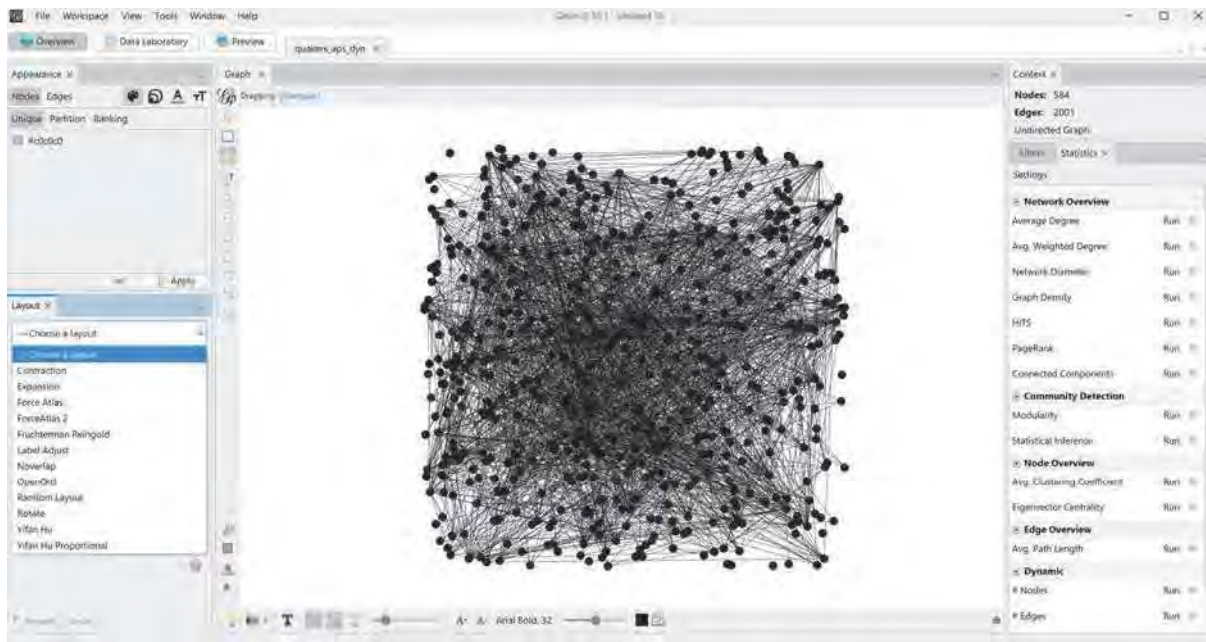


Figure 6.6 Gephi Visualisation Tool displaying the raw Project Seven data

- UBIRA for preservation and sharing of the dataset.⁵³⁸

Because the technology suite was complex (even though the data was seemingly not), a data pipeline and a data schema were thought essential, and they were designed ‘up front’ before any data handling work began. Datasheets⁵³⁹ were used to discipline the use of the third-party data that was donated and collected in CSV format from:

- Royal Anthropological Institute (metadata and microfilm).
- Wellcome Institute catalogues.
- *Protecting the Empire's Humanity: Thomas Hodgkin and British Colonial Activism 1830–1870* index (Laidlaw 2021).
- Friends House Archive manuscripts.

⁵³⁸ eData.bham.ac.uk (Accessed 12/12/2024)

⁵³⁹ <https://github.com/fau-masters-collected-works-cgarbin/datasheet-for-dataset-template/blob/master/datasheet-for-dataset-template.md> (Accessed 12/12/2024)

Project container structures were established at the outset. The organising principle of data organisation was that all of the data necessary for the P7 project would be held in containers (commonly folders in Explorer). In practice this meant that if a file needed to be used in several projects, then a separate copy would be placed in each project's container. This structure was necessary to enable GitHub, JNB and Gephi to function and it was universally adopted to efficiently manage navigation through hundreds of project files. Therefore, to avoid confusion, a container structure was adopted throughout the project (see Figure 6.7).

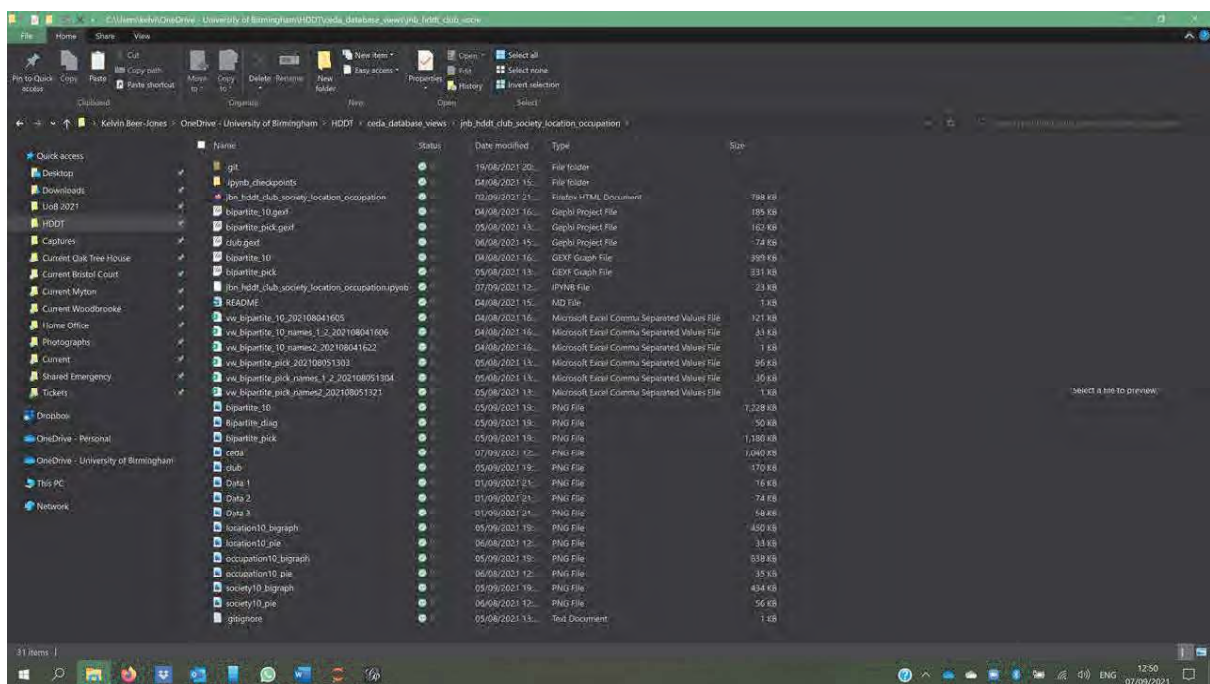


Figure 6.7 Example of a container for Project Seven.

6.18 Data collection, management and cleaning

All data, except for the RAI metadata, had full provenance tracking back to primary sources. Before cleaning and matching of donor datasets could take place, an Entity Relationship Diagram (ERD) was designed to hold the combined data in an SQL database (Figure 6.8). This is because data cleaning is never an object in itself, it is always directed to the need for the data to fit both the database and the data output requirements. A concern at this stage of any DH project is balancing (and trade-offs) between the integrity of the data itself and the needs of the database it will reside in and the other component HDDT technologies. It was discovered that cleaning exercises that made the data acceptable to SQLite were not totally acceptable to JNB, and then later Gephi, and so looping through data cleaning processes became the norm.

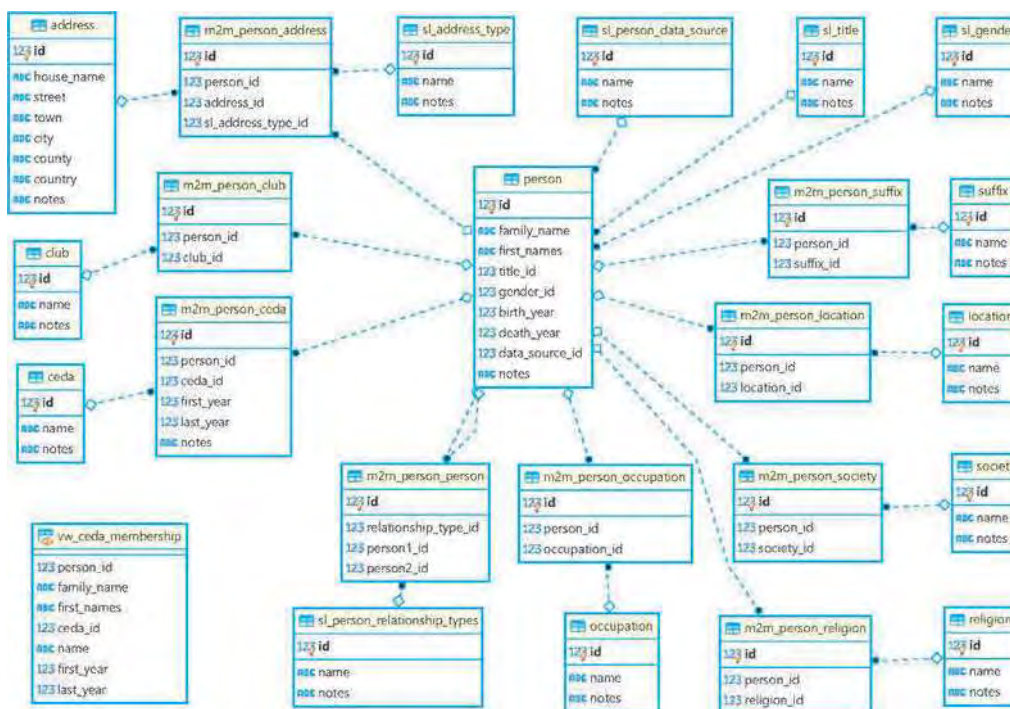


Figure 6.8 The HDDT ERD

6.18.1 Database specification

The database design had to address the strict data requirements of Gephi, because a key output from the HDDT is the production of NetworkX GexF files for Gephi to be able to visualise networks. This meant structuring data tables in the database with ‘nodes’ and ‘edges’ incorporated alongside primary fields so that later Gephi could recognise and interpret the data (Table 6.1). Data formatting and cleaning are always influenced by the technologies chosen for the study as well as having a ‘tidying up’ function.

Nodes (source and target)	Edges	Select list
person	m2m_person_person ²	y
CEDA	m2m_person_ceda	y
religion	m2m_person_religion	n
occupation	m2m_person_occupation	n
club	m2m_person_club	n
society	m2m_person_society	n
address	m2m_person_address	y
location	m2m_person_location	n
family	m2m_person_family	y
Attributes of person		
SI_gender		
SI_title		

Suffix (and m2m_person_suffix)

Table 6.1 Database specification

6.18.2 Donor data Excel sheets

A sample of each donor data source collected in Excel follows (Figures 6.9–6.11). This illustrates that although all of the datasets were in Excel, nonetheless the Excel sheets had no commonality in structure except that data was contained in rows and each row had person name as the key field. Because the database and its data are over forty years old, the data does not conform to any standards or conceptual models (see Chapter 4.2).

6.18.2.1 The RAI data (one record transposed)

Column head	Cell entry	Data cleaning challenges
name	Hodgkin	P7 Project Authority Index component. Action: preserve
FIRST NAME	Thomas (1)	Unconventional name entry. This is a P7 Project Authority Index component. Action: preserve
TITLE		Can be multiple titles recorded as a single string. Freeform entries (Captain, Capt, Capt)
suffix	MD, FRGS	Single string
gender	M	Often no entry
address	9 Lower Brook Street [1838] 35 Bedford Sq	Occasional carriage returns
location	London	Extraneous characters. Action: strip out
society	ESL APS	Single string, carriage returns
elected ESL	1844.02.01	Consistent entries
elected ASL		Consistent entries
elected AI		Consistent entries
Elected APS	1838.05.16	Consistent entries
elected LAS		Consistent entries (always blank)

Figure 6.9 RAI data Thomas Hodgkin1 part 1

membership	ESL Ordinary Fellow	Single string. Note
	APS Ordinary fellow [Life Member]	capitalisation of 'fellow'
		Action: standardise capitalisation
office notes	APS Council 1838 member APS Council 1839 member APS Council 1840 member	Single string.
	APS Council 1847 member and Secretary	Note: blank entries
	ESL Council 1844 Member ESL Council 1844-45 Member ESL Council 1845-46 Member ESL Council 1846-47 Member ESL Council 1847-48 Member ESL Council 1848-49 Member ESL Council 1849-50 Member	Note: additional comments
	ESL Council 1850-51 Vice President ESL Council 1851-52 Vice President ESL Council 1852-53 Vice President ESL Council 1853-54 Member	
	ESL Council 1854-55 Member	
	ESL Council 1855-56 Vice President [replaces Greenough deceased] ESL Council 1856-57 Vice President	
	ESL Council 1857-58 Member	
	ESL Council 1858-59 Member ESL Council 1859-60 Member ESL Council 1860-61 Member ESL Council 1861-62 Member ESL Council 1862-63 Member ESL Council 1863-64 Member ESL Council 1864-65 Member	
	ESL Council 1865-66 Member [dies Apr. 1866]	
house notes	Founder member ESL. 1844.01.02 becomes trustee	Single string
	number after name to distinguish from another with same name	Note: typing errors, spelling mistakes. Missing 'S' in Sub
	Trustee Jan 1844; Publications Committee Apr 1844; Library Committee Apr 1844; House Committee Apr 1844; Committee to consider future of Society Apr 1849; Committee to consider means of improving prospects of Society Mar 1850	Occasional semi colons Varied date forms
	Trustee Jun 1850; Committee to draw up an address to the Society Jun 1850; Sub- Committee on printing Jun 1850; Committee to revise Rules and Regulations Sep 1850; delegate to BAAS at Ipswich 1851; sub-committee to consider papers 1852; sub-committee to enquire for accommodation Jul 1852; sub-committee to report on another volume 1853; ub- committee to examine Dr King's claim, Oct 1853; Finance Committee 1856; Committee to consider advancing the Society 1857; Committee for selection and publication of papers Feb 58; Committee for preparing the house list May 58; Publication committee Jul 58; Committee on changes in Council May 60; sub-committee to report on Mr Clarke's portraits Mar 62	
left	1862 last listed	Varied date forms Added text
references	A1; A2:1; JES 1; EJ 69; TES vol III 65	Mixed internal references and findable references
	ESL list 1862	
	APS A111/1 First Annual report May 16th 1838; Second Annual report May 21st 1839; Third Annual report 23rd June 1840; Tenth	
	Annual report 17th May 1847	
	A111/3 The Colonial Intelligencer 1847- 1848	
Occupation	medical	Extraneous added characters
dates	1798	Unexplained entries
death	1866	Variable date forms

Figure 6.10 RAI data Thomas Hodgkin1 part 2

Notes- from elsewhere	Thomas Hodgkin (1798–1866) was an English physician, considered one of the most prominent pathologists of his time and a pioneer in preventive medicine. He is now best known for the first account of Hodgkin's disease, a form of lymphoma and blood disease, in 1832. Hodgkin's work marked the beginning of times when a pathologist was actively involved in the clinical process. He was a contemporary of Thomas Addison and Richard Bright at Guy's Hospital.	Unattributed text
Publications	<p>Hodgkin published as a book his Lectures on Morbid Anatomy in 1836 and 1840. His major contribution to the teaching of pathology, however, was made in 1829, with his two volumed work entitled The Morbid Anatomy of Serous and Mucous Membranes, which became a classic in modern pathology.</p> <p>Hodgkin was one of the earliest defenders of preventive medicine, having published On the Means of Promoting and Preserving Health in book form in 1841. Among other early observations were the first description of acute appendicitis, of the biconcave format of red blood cells and the striation of muscle fibers.</p> <p>Hodgkin also translated with Thomas Fisher, from the French of William-Frédéric Edwards, On the Influence of Physical Agents on Life (London, 1832; Philadelphia 1838).[41][42] Edwards was a vitalist in physiology, who studied the effect of physical forces on processes in living organisms.[43] The work as it appeared in</p>	Unattributed text. UK and US spelling
	English was much more than a translation, since it contained an appendix of over 200 pages containing two dozen papers, a compendium of medicine and science tangentially related to themes in Edwards, but related to his general approach. It included early work by Hodgkin and collaboration with Lister, as well as something on electricity and meteorology.[44] He also published The Means of Promoting and Preserving Health (London, 1840), of which a second edition appeared in 1841, and an Address on Medical Reform (1847).	
house publications	<p>The progress of ethnology On the Guanches. Printed Jun 1845</p> <p>On the ancient inhabitants of the Canary Islands. Read 21 May 1845</p> <p>Obituary notice of Dr Prichard. Read 28 Feb. 1849. Printed</p> <p>communicates On an Indian tribe in the NW of America by the late Manuel Cardenas. 8 Dec 1852</p> <p>On some Darien Indians by Manuel Cardenas. Read by Hodgkin 12 Jan 1853 On the Bedouins</p>	Unattributed text
clubs		Extraneous characters
societies	Royal Geographical SocietyPhilological SocietySociety for the Diffusion of Useful KnowledgeBritish Association	Single string Extraneous characters
RAI material details		
Related material details	Wellcome Library: papers	Imprecise attribution
	(see also A51/12/7 .1 Miss M.M. Scaife to FS, 30 Oct. 1944 - sends reprint of Dr James Hunt's 1865 address from her step- grandfather, Dr Thomas Hodgkin's papers which she is destroying (autogr.)	Note: destroyed primary source.
		Note: unconventional abbreviation
	.2 FS to Miss Scaife, 2 Nov. - very pleased to receive reprint; would welcome any other relevant papers from Dr Hodgkin's estate	Note: unidentified text (tpc.)
	(tpc.); no manuscripts by Dr Hodgkin were received	

Figure 6.11 RAI data Thomas Hodgkin1 part 3

6.18.2.2 APS data

Data was collected onto the Excel workbook by hand, by reading through eight rolls of microfilm at the RAI archives consisting of the publications of the APS, one worksheet for each of over forty observations (Figure 6.12).

1838	First Annual Report of the Aborigines Protection Society			
	Exeter hall	16/05/1838		
	Paper held in the box containing the A111 collection of 8 reels microfiche			
	T Fowell	Buxton	President	
	Committee			
	William	Allen		
	G F	Angas		
	William	Aldam		
	E	Baines	MP	
	E N	Buxton		
	S	Bannister		
	Augustus	D'Este	Sir	Bart
	Josiah	Forster		
	S	Gurney	Jun	
	C	Hindley	MP	
	Gurney	Hoare		
	Robert	Howard		
	W M	Higgins		
	T	Hodgkin	MD	
	M	Hutchinson	Jun	
	A	Johnston		
	R	King		
	S	Lushington	Dr	MP
	C	Lushington	MP	
	J	Pease	MP	
	T	Roscoe		
	Culling Eardley	Smith	Sir Bt	
	Ebenezer	Smith		
	Hull	Terrell		
	A	Wells	Rev	
	S	Wilkin		
	H	Tuckett	Treasurer	
	J J	Freeman	Honorary Secretary	
	J H	Tredgold	Honorary Secretary	

Figure 6.12 Sample APS data collection in Excel

6.18.2.3 QCA data

Information was copied by hand by reading manuscripts at Friends House Quaker Archives, London (Figure 6.13).

Committee on the Aborigines			
(Mfs 7th of the 7th Month 1837 (p419-420))			
William Allen			
Thomas Christy			
John Hodgkin Jun			
John Thomas Barry			
William Forster			
Robert Forster			
Abram Rawlinson Barclay			
George Stacey			
Robert Howard			
Henry Knight Jun			
Josiah Forster			
William Hargreave			
Samuel Darton			
John Sanderson			
Peter Bedford			
Thomas Norton Jun			
Richard Barrett			
John Hamilton			
Edwd Harris			
Robert Alsop Jun			
John Kitching			
Joseph Storrs			
Geo Holmes			
Joseph Shewell			
John Bell			
Joseph Neatby			
John Barclay			
Joseph Talwin Foster			

Figure 6.13 Quaker Committee on the Aborigines (four Excel workbook sheets)

6.18.2.4 Quakers data

Information was extracted from a personal Gedcom file by the Secretary of the Quaker Family History Society (Figure 6.14).

family_name	first_names	title	aps_first_year	aps_last_year	qca_first_year	qca_last_year	religion	notes	
Barclay	Abram Rawlinson				1837	1839	Quaker	Abraham Rawlinson Barclay (1793-1845), author, s. of Robert Backlay (1758-1816)	
Barclay	Eliza	Mrs	1851	1866				Eliza (Backhouse) Barclay (1812-1884), of Darlington; 27-page obit. in 1885 Annual Monitor	
Barclay	G		1856	1856				Probably (Joseph) Gurney Barclay, FRAS (1816-1898), banker, of Leyton, Essex; 5-page obit. in 1900 Annual Monitor	
Barclay	J G		1851	1864				(Joseph) Gurney Barclay, FRAS (1816-1898), banker, of Leyton, Essex; 5-page obit. in 1900 Annual Monitor	
Barclay	J Gurney		1852	1852				(Joseph) Gurney Barclay, FRAS (1816-1898), banker, of Leyton, Essex; 5-page obit. in 1900 Annual Monitor	
Barclay	John				1837	1839	Quaker	John Barclay (1821-1889), druggist, of Falmouth	
Barclay	Joseph G		1865	1867				(Joseph) Gurney Barclay, FRAS (1816-1898), banker, of Leyton, Essex; 5-page obit. in 1900 Annual Monitor	
Barclay	Robert	Late	1838	1855				Probably Robert Barclay (1787-1853), London banker	
Barclay	x	Mrs	1853	1855				Insufficient identifying evidence	
Barnes	x		1838	1840				Insufficient identifying evidence	
Barrett	Jonathan		1860	1861				Jonathan Barrett (1790 - after 1847), London brass-founder	
Barrett	R		1853	1860				Richard Barrett (1784-1855), London brass-founder; 5-page obit. in Annual Monitor	
Barrett	Richard				1837	1839	Quaker	Richard Barrett (1784-1855), London brass-founder; 5-page obit. in Annual Monitor	
Barrington	Richard		1864	1867				Probably Richard Barrington (c. 1797-1890), of Monkstown, Dublin; death announced in Annual Monitor	
Barrow	R C		1860	1864				Richard Cadbury Barrow (1827-1894), tea & coffee merchant, of Birmingham	
Barry	J T		1851	1863				John Thomas Barry (1789-1864), pharmaceutical chemist, of London	
Barry	John Thomas		1852	1853	1837	1847	Quaker	John Thomas Barry (1789-1864), pharmaceutical chemist, of London	

Figure 6.14 Quakers

6.18.2.5 Quaker families data

Information was extracted from a personal Gedcom file by the Genealogist (Figure 6.15).

Kelvin's ID	Ben's ID	Family nam	First name	School	Occupation	Related to	Related to	Related to	Notes	Updated notes
		1 Albright	Arthur	Friends' sch	printer and	2,3,4			Possibly the Quaker Arthur Albright (1811-1900) who has an entry in the Oxford DNB	
		2 Albright	John M		draper groc	1,3,4			John Marshall Albright, Quaker (1815-1909), Charlbury; 13-page obit. in 1910 Annual Monitor	
		3 Albright	Rachel			1,2,4			Could be either Rachel Albright (1818-1906), sister of John M. Albright; or their mother Rachel Albright, née Tanner (1776-1867)	sister of 1, 2, and probably 4 - assuming this is the younger Rachel
		4 Albright	William		grocer, dra	1,2,3			I have no plain William Albright who fits, though there was a William Whitlark Albright (1819-1864), of Lancaster and Sheffield - It would seem odd, though, that his middle name would have been omitted	probably William Tanner Albright, (c. 1805 - ?), brother of 1, 2, & 3
23		5 Aldam	William		stuff merch		6	236	Probably William Pease (later Aldam - changed by deed poll) (1779-1855), of Darlington	
		6 Aldam	William Jun		magistrate,		5	236	Probably William Aldam (1813-1890), MP for Leeds, son of the earlier William Pease	
		7 Alexander	Frederick			8,9		10	Two Quakers of this name - so need more identifying detail	Frederick Alexander (1811-1893); identification deemed probable, in light of relationships
		8 Alexander	G W		banker & pl	7,9		10	George William Alexander (1802-1890) - entry in Oxford DNB	
		9 Alexander	Henry		ironmonger	7,8		10	Henry Alexander (1808-1884)	
		10 Alexander	R D		banker		#####	7,8,9,132	Richard Dykes Alexander (1788-1866), Ipswich	
		11 Allen	Samuel		brewer & m		12	284	Possibly Samuel Allen (1771-1868), of Witham and Hitchin; 4-page obit. in 1870 Annual Monitor	
27		12 Allen	Stafford		millier & ma		11	#####	Stafford Allen (1806-1889), of London; 9-page obit. in 1890 Annual Monitor	
		13 Allis	Thomas		naturalist, superintend		#####	#####	Thomas Allis (1788-1875); entry in Milligan's Biographical Dictionary of British Quakers in Commerce and Industry	
		14 Alsop	Robert						Probably the same as Robert Jun., who was the son of a Robert, but the father d. in 1850	probably duplicate
		15 Alsop	Robert Jun						Robert Alsop (1803-1876), pharmaceutical chemist, of London; 6-page memoir in 1877 Annual Monitor	no relationships found
		16 Appleton	John D	Ackworth					John David Appleton (1830-1907), 2-page obit. in Ackworth Old Scholars' Assn Annual Report	no relationships found
		17 Arch	John					85	There were Quakers of this name, but the closest match died in 1853	relationship is to the closest match

Figure 6.15 Quaker family relationships

6.18.3 Data cleaning

A fundamental concept of data cleaning and handling is that data is not simply overwritten by the researcher if it is felt to be mistaken, inaccurate or poorly constructed. The only exception to this rule in the P7 project was to enforce the PROPER capitalisation of person names (because they are both the project primary key and also the project Authority Index, necessitating a minimum of conformity). The RAI data required considerable cleaning. Mine and The Genealogist's data were relatively much cleaner (this is because this data was recently created specifically for the P7 project, whereas the RAI metadata was over forty years old and deliberately styled as a finding aid).

A three-dimensional Excel workbook methodology was used for data cleaning. Sheet 1 was the RAI dataset, sheet 2 a copy of the RAI dataset and sheet 3 an audit of sheets 1 and 2, where TRUE was returned if the values in the same cell in sheets 1 and 2 were the same, and FALSE if they were not. Unexpected FALSE values alerted the researcher to errors in data cleaning process. Sheet 1 then became the data cleaning sheet, working on one column at a time by inserting a column next to the column to be cleaned and using Excel data cleaning formulas to unpack and represent the data, such as:

- =PROPER to impose proper case capitalisation
- TRIM() – to delete excess spaces
- =CLEAN(text) – to delete non-printing characters
- =LEFT(A2, SEARCH("-",A2,1)-1) – to search for specific text in a string
- =MID(A2, SEARCH("-",A2) + 1, SEARCH("-",A2,SEARCH("-",A2)+1) - SEARCH("-",A2) - 1) – to extract text from the middle of a string
- =MID(A2, SEARCH(CHAR(10),A2) + 1, SEARCH(CHAR(10),A2,SEARCH(CHAR(10),A2)+1) - SEARCH(CHAR(10),A2) - 1) – to remove line breaks

- `=RIGHT(A2,LEN(A2) - SEARCH("-", A2, SEARCH("-", A2) + 1))` – to split the string from the right

After cleaning data in the target column and the audit sheet 3 showing that no data in any other cells had changed, the file was then saved as a CSV file (which saves only sheet 1). VSC was then used to upload the file to GitHub for version control. This CSV file became the stage 1 cleaning file. To clean the next column a new three-dimensional Excel workbook was then set up with sheets 1 and 2 this time being the stage 1 cleaning file and the cleaning took place on the next column of data. The process was repeated fourteen times until all columns were considered clean. GitHub was used for version control because it highlighted each cell value that had changed at each stage of the cleaning process. This allowed visual auditing to take place.

The rendering of names imported from CSV in the SQL database sometimes revealed additional differences that were not visible in Excel. About 100 records failed later when uploaded to the SQL database, another 50 failed when data passed through the technology pipeline from SQL to JNB, and a further 20 failed when data passed from JNB to Gephi. These errors were corrected in the database using DBeaver.

The data cleaning process followed a strict methodology.

- Excel functionality together with VSC and GitHub for version control allowed for auditing and memorialising of the data cleaning activity. A prepared script was used to compile a set of Excel spreadsheets in strict sequence (01, 02, 03...14, one

spreadsheet for each step in the cleaning sequence). Then a set of CSV copies was made of each of the XLS files generated in the cleaning exercise.⁵⁴⁰

- Visual Studio Code was used to send the resulting CSV files in strict sequence to GitHub to facilitate the production of a version control record of the entire cleaning exercise and to fully memorialise all of the data cleaning actions that had taken place in it.
- Testing that the final project combined CSV dataset derived from the XLS donor data set would then produce the required project database in SQLite.

6.18.4 Data cleaning reconciliation

A manual record was made of all expected deletions and modifications to the database, which was updated after each of six modification scripts were run on the SQLite database (see Figure 6.16).

⁵⁴⁰ See Project Seven Report 7.3.8 for a sample script used to clean APS data.

	Data clean	KBJ files						CEDA Database			
RAI dataset		2260						Datasource = 1	1988	272	
KBJ dataset		1159									
QCA			35					Datasource = 2	16	19	
APS			1201					Datasource = 3	1090	111	
Already in RAI (update only)			-77							-77	
New Records			1159						3094		
Total donor records		3419									
Start date > 1871 delete record		-51									
Make last_Year = first_year where last_year = NULL		-183									
		183									
Make last_Year = first_year where last_year > 1871		-1451									
		1451									
first_names NULL, replace with 'x'		-129									
		129									
first_names 'X', replace with 'x'		-4									
		4									
Duplicate m2m_person_ceda records		-10									
		10									
Delete persons not in any ceda		-170									
Delete duplicate persons found by Ben beck		-50									
Delete duplicate persons found by Beer-Jones		-105									
Total added or deleted		-325									
ceda database 2020/07/16 by calc		3094									
Current database		3094									
gain / loss		0									

Figure 6.16 Data cleaning reconciliation

6.18.5 RAI data challenges

A visual examination of the RAI data revealed issues that would need to be addressed in data cleaning:⁵⁴¹

- Column heads (and other entries) were sometimes all lower or all upper case, sometimes with common initialisation and often with a mixture of capitalisation.

This can occur if the donor database model does not have rules about data styling.

⁵⁴¹ See Project Seven Report 7.3.3.

- Unconventional name entries (e.g. Thomas Hodgkin1).⁵⁴²
- Multiple surnames recorded as a single string. Freeform entries were common (Captain, Capt., Capt).
- Multiple entries entered as a single string (e.g. suffixes, for example MD,FRGS or MD-FRGS or MD.FRGS).
- Unexpected carriage returns.
- Extraneous characters not visible in Excel.
- Typing errors, spelling mistakes, US and UK spelling, varied date forms.
- Imprecise references, unexpected abbreviations, unidentifiable text.⁵⁴³

6.18.6 P7 author data challenges

The challenges here were in copying EBP data by hand from microfilm where data was occasionally out of focus, poorly photographed (dark or blurred image) and sometimes parts of pages were missing. The microfilm was therefore sometimes difficult to read.

Fortunately, the object was to collect person names and these were often more carefully written/typed out than the accompanying text (see Figure 6.17).

⁵⁴² The data file cell for House Notes records number after name to distinguish from another with same name. See Project Seven Report Table 7.3.8.6 The RAI data structure.

⁵⁴³ See Project Seven Report Table 7.3.8.6, The RAI data structure.

ABORIGINES PROTECTION SOCIETY.

3

President.

T. FOWELL BUXTON, Esq.

Committee.

WILLIAM ALLEN, Esq.
G. F. ANGAS, Esq.
WILLIAM ALDAM, Jun. Esq.
E. BAINES, Esq. M.P.
E. N. BUXTON, Esq.
S. BANNISTER, Esq.
SIR AUGUSTUS D'ESTE, Bart.
JOSIAH FORSTER, Esq.
S. GURNEY, Jun. Esq.
C. HINDLEY, Esq. M.P.
GURNEY HOARE, Esq.
ROBERT HOWARD, Esq.
W. M. HIGGINS, Esq.
T. HODGKIN, Esq. M.D.

M. HUTCHINSON, Jun. Esq.
A. JOHNSTON, Esq.
R. KING, Esq.
Dr. S. LUSHINGTON, Esq. M.P.
C. LUSHINGTON, Esq. M.P.
J. PEASE, Esq. M.P.
T. ROSCOE, Esq.
H. E. RUTHERFOORD, Esq.
SIR CULLING EARDLEY SMITH, Bt.
EBENEZER SMITH, Esq.
HULL TERRELL, Esq.
Rev. A. WELLS.
S. WILKIN, Esq.

Treasurer.

HENRY TUCKETT, Esq.

Honorary Secretaries.

Rev. J. J. FREEMAN, Walthamstow.
J. H. TREDGOLD, Esq. 41, Wellclose Square.

The object of this Society is to assist in protecting the defenceless, and promoting the advancement of uncivilized Tribes.

A Subscription of One Guinea a year, or a Donation of Ten Pounds constitutes a Member.

The Society is desirous of promoting the formation of Auxiliary Associations, both at home and abroad.

Subscriptions or Donations in aid of the Funds of the Society will be thankfully received by the Treasurer, the Secretaries, or any Member of the Committee.

Offices, No. 4, BLOMFIELD STREET, FINSBURY.

Figure 6.17 Image of the First Annual report of the (Aborigines Protection Society 1838)

The data challenges of reading manuscript records in Friends House Quaker archives to extract the names of the committees that formed the QCA were less problematic because the primary sources themselves could be consulted. It helps if the reader is practised in reading nineteenth-century cursive script (see Figure 6.18).

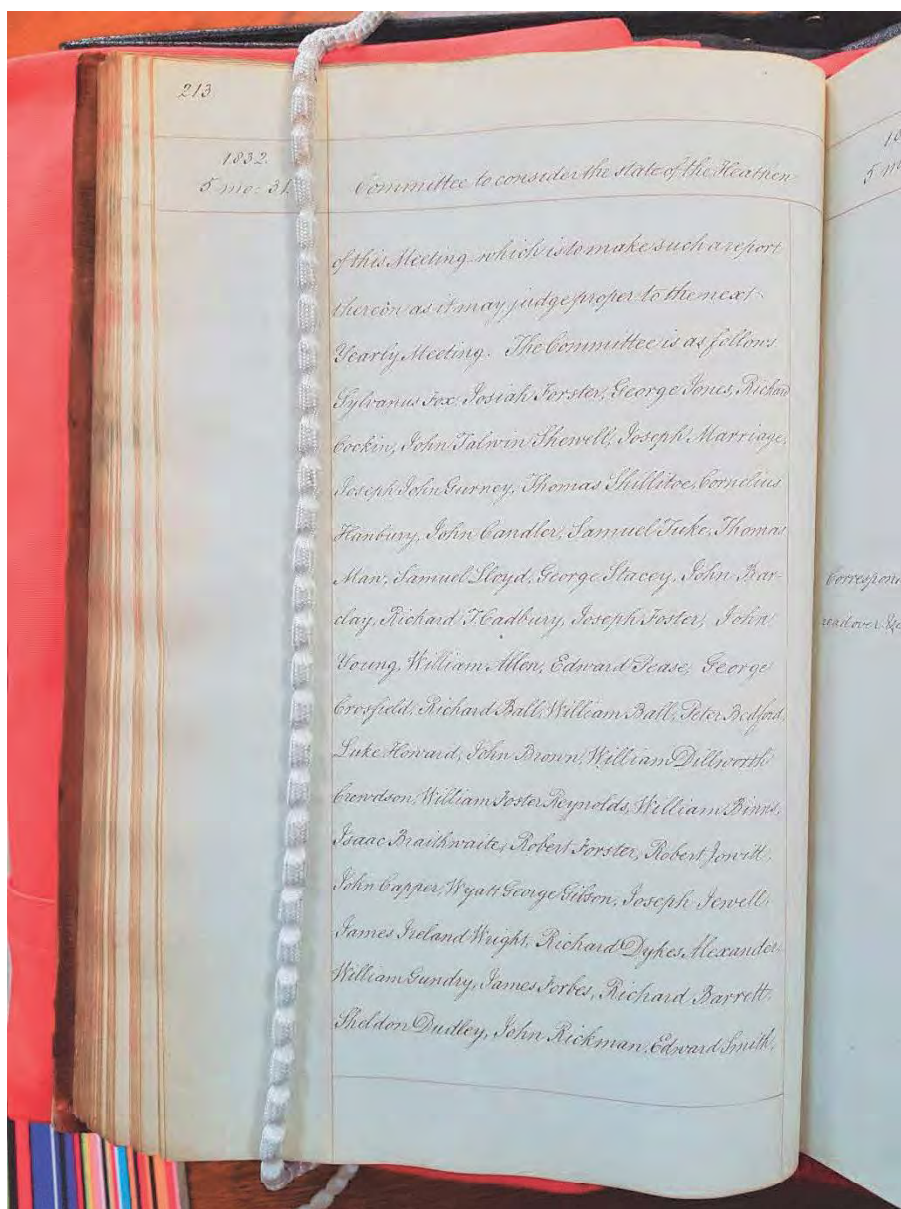


Figure 6.18 Image of London Yearly Meeting 1832 – the names of members of the Quaker Committee on the State of the Heathen (part of the QCA)

6.18.7 The genealogical data challenges

There were only two challenges, both relating to the family relationships data, where the Genealogist drew on his previous work identifying Quakers in the database:

- The donor spreadsheet did not include data for column 'A'. This was obtainable from the previous year's donor data offering, which then had an allocation of 'Kelvin's ID' numbers when that data was uploaded into the HDDT. This new donor spreadsheet was originally generated from a View of 'religion = quaker' taken from the updated HDDT (and therefore based on the first of the Genealogist's donor data offering). It was possible that some data might have changed in the processing so far and so a match of 'Kelvin's ID2' to the genealogists UID - 'Ben's ID', required close scrutiny.
- Although great care had been taken by the genealogist to ensure that data in the key columns 'G =Related to (category 1), H = Related to (category 2), I = Related to (category 3)' was consistent, Excel formatting inconsistencies were found. These had to be resolved in data cleaning to avoid the misidentification of data later (cells formatted as numbers must be re-designated as 'general' and the data re-entered). For an example see cell G26, where persons 173 and 174 were entered as a single whole number 173174 with Excel formatting rendering the number as 173,174 (using the US comma convention for representing large numbers), whereas the desired entry should have been 173,174 with a non-numeric comma used to separate the numbers 173 and 174. For this to work in Excel the cell must be set to 'general' and not 'number'.

The genealogist provided notes, which included a challenge to the acceptance of a few records, and these notes were taken into consideration.

6.18.8 Data matching

The P7 project accepted the RAI person names as the project authority index. Record matching therefore had to take place only to merge the APS data. Discussions with the RAI archivist at project commencement confirmed that the RAI generally accepts the APS as a foundation society and wanted this project to include the APS as members of the CEDA. However, several APS members to be added were already present in the RAI data file because they were members of the ESL, ASL or AI. The APS names extracted from the microfilm publications of the APS contain only names, with no attributes, unlike the RAI data which has many attributes. It is likely that attributes were not necessary to the APS membership secretary because the members knew each other personally and on each occasion names were probably collected in person by one individual (the membership secretary usually being present at each annual meeting). Similarly, the names collected from the Friends House archive manuscripts were all known to the meeting members compiling the minute. This familiarity between members is also likely within the ESL, ASL and AI membership lists, at least in the early years.

RAI attribute data was reported as sourced by several archivists a long time ago and from a range of sources collected over many years. Name recording styles varied from year to year throughout (possibly because different officers collected names and followed different conventions). However, the total number of names considered for matching in the whole

CEDA dataset was relatively small and so mismatching of names, while always possible, was judged to be infrequent and not likely to unduly influence the P7 project outcomes.

Only fifty-two person records were deleted as duplicates in the name matching process.

Good practice in EBP requires that care be taken when matching names with decisions based on probability. In this case it was observed that, when possible name duplication arose, it was usually because two members with the same name were present at a meeting (e.g. father and son). Usually the membership secretary differentiated these names using 'Junior' or 'Senior'. The handing down of names through the generations was a common practice in the nineteenth century and especially among Quakers. In the RAI data set the archivist differentiated between Thomas Hodgkin MD and his nephew Thomas Hodgkin by appending the digit '1' to Thomas Hodgkin MD's name. This is what the 'based' in EBP means – a judgement must be made by the researcher based on the evidence available. Any decisions made by a researcher are based on probability and observance; they are therefore subject to challenge. The object in building an HDDT and a dataset is to welcome challenge.

This was particularly the case when identifying Quakers among the person file entries. Data on membership dates in the RAI data was always available (because that was the object of the RAI database) and frequently the information in the Office Notes and House Notes fields could be used to corroborate the information provided by the Genealogist in his family relationships file. Nevertheless, many potential Quakers were rejected as Quakers because the Genealogist made clear that uncertainty arose in matching. They remain in the database without an attribute 'religion'. Overall this means that some members of the APS were added to the database as new persons when they might already be present in the database due to their pre-existing ESL, APS or AI memberships. The total number of members of the

CEDA may be over-recorded by less than 100. On the other hand, Quakers present in the database may have been under-counted. This is because there was insufficient evidence to reasonably identify them. The number of Quakers not identified in the database may be between 100 and 200.

6.18.9 VSC and GitHub

Once the cleaning of data in the donor Excel spreadsheet was complete it was necessary to convert the Excel files to CSV files to render them in a format suitable for GitHub. Each CSV file (beginning with the unmodified version) was then committed to GitHub to facilitate version control and to make an open-source memorialisation of the entire data cleaning process.

6.19 Data analysis and visualisation⁵⁴⁴

6.19.1 Members of the CEDA

1. Quaker Committee on the Aborigines (QCA), 31 members.
2. Aborigines Protection Society (APS), 1171 members.
3. Ethnological Society of London (ESL), 748 members.
4. Anthropological Society of London (ASL), 1334 members.

⁵⁴⁴ The database structure reflects the ERD (see Project Seven Report Figure 7.4.1).

5. Anthropological Institute (AI), 610 members.

See Figure 6.19.

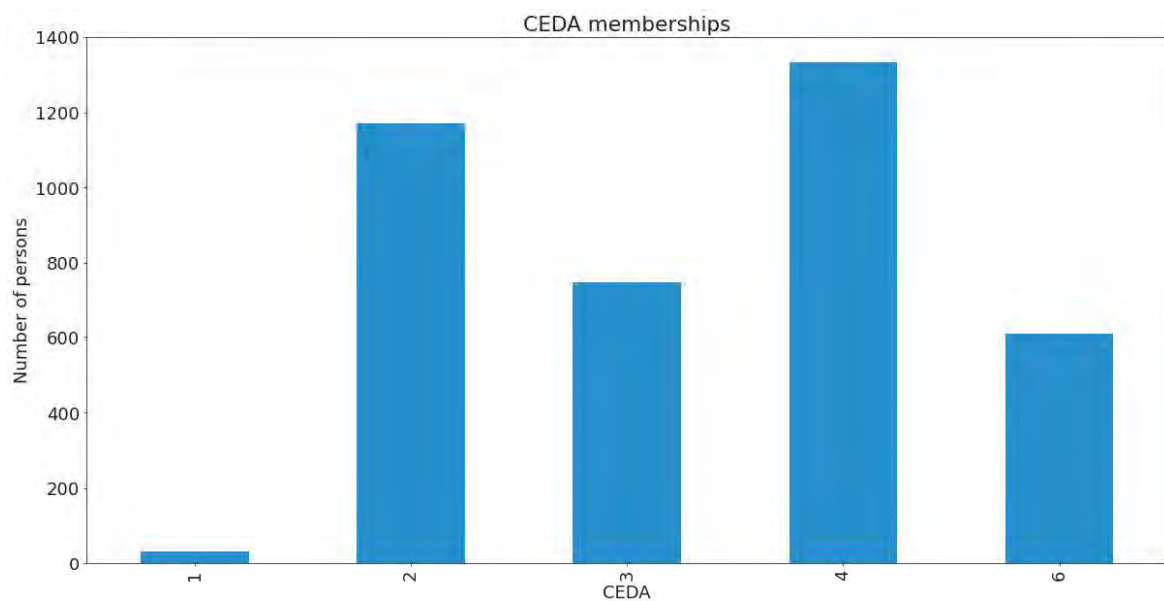


Figure 6.19 CEDA memberships by society

6.19.2 Attributes tables

- ceda (6)
- club (68)
- location (83)
- occupation (93)
- person (3095)
- suffix (155)
- religion (4) – only one group is present, 1 = Quaker
- society (260)

6.19.3 M2M relationships

- m2m_person_ceda (3894)
- m2m_person_club (323)
- m2m_person_location (2061)
- m2m_person_occupation (1883)
- m2m_person_person (2099)
- m2m_person_religion (593)
- m2m_person_society (1238)
- m2m_person_suffix (1351)

6.19.4 The person file

The person file is at the heart of the database and the eight attribute files are linked to it with 'many to many' possible relationship linkages between the files (see Figure 6.20).

	id	family_name	first_names	title	gender_id	birth_year	death_year	data_source_id	notes
1	1	A Beckett	Arthur William	(NULL)	1	1,844	1,909	1	17 King Street, S. James's, S.W. 1
2	3	Adam	Andrew Mercer	(NULL)	1	(NULL)	(NULL)	1	Boston, Lincolnshire
3	4	Adam	H R	(NULL)	1	(NULL)	(NULL)	1	Old Calabar, W. Africa
4	5	Adam	William	(NULL)	1	(NULL)	(NULL)	1	(NULL)
5	6	Adams	Henry John	(NULL)	1	(NULL)	(NULL)	1	14 Thornhill Square, N.
6	7	Adams	William (1)	(NULL)	1	(NULL)	(NULL)	1	(NULL)
7	8	Adams	William (2)	(NULL)	1	1,820	1,900	1	5 Henrietta Street, Cavendish Sq
8	11	Adlam	William	(NULL)	1	(NULL)	(NULL)	1	9 Brook Street, Bath [1863] Marv
9	12	Agassiz	Louis	(NULL)	1	1,807	1,873	1	Cambridge, Mass
10	13	Agathides	Anastasius	(NULL)	1	1,805	1,881	1	26 Kildare Terr. Westbourne Par
11	14	Aglio	Augustine	(NULL)	1	1,777	1,857	1	(NULL)
12	15	Agnew	Joseph	(NULL)	1	(NULL)	(NULL)	1	47 Bath St., Glasgow
13	16	Ainsworth	William Francis Harrison	(NULL)	1	1,807	1,896	1	Ravenscourt Villa Hammersmith
14	17	Airston	William Baird	(NULL)	1	(NULL)	(NULL)	1	5, Andrew's, Fife
15	18	Aitken	Alexander Muirhead	(NULL)	1	(NULL)	(NULL)	1	13 St George's Terrace South Ke
16	19	Aitken	Thomas	(NULL)	1	(NULL)	(NULL)	1	District Lunatic Asylum, Inverne
17	22	Alcock	Rutherford	His Excellency Sir	1	1,809	1,897	1	Japan [1862] China [1868]
18	23	Aldam	William	(NULL)	1	1,813	1,890	1	(NULL)
19	25	Aley	Frederick W.	(NULL)	1	(NULL)	(NULL)	1	8 Thurloe Place, South Kensingt
20	26	Allan	J. McGregor	(NULL)	1	(NULL)	(NULL)	1	26 Park St., Southampton St., Ca
21	27	Allen	S. Stafford	(NULL)	1	1,840	1,870	1	2 Paradise Row, Stoke Newingto
22	28	Allen	William (Capt.)	Capt.	1	(NULL)	(NULL)	1	(NULL)
23	29	Allin	George	(NULL)	1	(NULL)	(NULL)	1	14 High Street, St Albans Lancin
24	30	Alpe	Charles Hamond	(NULL)	1	1,857	1,882	1	37 Montpelier Rd. Peckham [18
25	31	Alston	Crewe	(NULL)	1	1,828	1,901	1	38 Belkize Park, Hampstead
26	33	Amhurst	William Amhurst Tyssen	RT. Hon. Lord	1	1,835	1,909	1	Didlington, Brandon, Norfolk [1
27	34	Amner	George	(NULL)	1	1,873	1,901	1	Reading, Berkshire
28	35	Anderson	Edward C	(NULL)	1	(NULL)	(NULL)	1	3 Tavistock Street, Covent Gard
29	36	Anderson	John	(NULL)	1	(NULL)	(NULL)	1	Ulverstone
30	37	Anderson	Joseph	(NULL)	1	1,832	1,916	1	Wick, Calthness
31	38	Andrew	W. P.	(NULL)	1	(NULL)	(NULL)	1	(NULL)
32	39	Andrews	x	(NULL)	1	(NULL)	(NULL)	1	5 White Hart Court
33	40	Anketell	Matthew John	(NULL)	1	(NULL)	(NULL)	1	30 Downshire Hill, Hampstead 9

Figure 6.20 The person file

6.19.5 Select queries

Further data cleaning was needed when the data could be managed and viewed more easily in the database using DBeaver. This was done by using select queries; for instance select queries identified that several records contained dates that were not credible. By referring back to the RAI metadata, it was possible to resolve the problem and then correct the dates in the database (see Figure 6.21). Similar data cleaning actions took place when the Quaker records were added to the database – this was the last dataset added (see Figure 6.22).

```

-- Modifications based on http://localhost:8888/notebooks/DataShare/dhdt_projects/test_area/8_Mike_training_time_series/time_series_test.ipynb

-- 1. DataFrame(.info)
-- birth_year shows as 'float' s/be 'int64', Possible negative birth year in dataset?
-- age_first_year shows as 'float' s/be 'int64.' Possible removed when birth_year cleaned?
-- Data cleaning of records required:

● --
-- person_id years_member
--1258 -1
--1266 -1
--1519 -1
--590 -1
--1355 -1
--1725 -1
--1895 -1680
--643 -2
--746 -7
--1160 -7
--1268 -1
--1483 -1
--1740 -2
--1991 -1
--2024 -1
--2447 -21

● --1258 -1
Select *
From vw_ceda_membership_gephi vcmg
Where id = 1258;

● Update m2m_person_ceda
Set first_year = '1862'
Where person_id = 1258;

● --1266 -1
Select *
From vw_ceda_membership_gephi vcmg
Where id = 1266;

● Update m2m_person_ceda
Set first_year = '1869'
Where person_id = 1266;

```

Figure 6.21 Correcting date errors

```

-- Changes to m2m_person_religion table from QFHS data on relationships which identified persons who are not Quaker
-- Start = 649 Quakers - 6 Not Quaker - 50 duplicates = 593 Quakers in the database after this exercise (check)
-- 6 persons are not Quaker (but they remain as persons):
-- Delete duplicate records (The database holds undirected relationships)
-- and so a record where Jack is related to John is not required if there is another record: John is related to Jack)
-- 585 Null records were present in the m2m_person_religion table
--post exercise stats:

SELECT count(*) FROM person;
--answer = 3094 (was 3095, then minus William allen, see last code block)

SELECT count(*) FROM m2m_person_address;
--answer = 0 (no change)

SELECT count(*) FROM m2m_person_ceda;
--answer = 3983 (was 3982)

SELECT count(*) FROM m2m_person_club;
--answer = 323 (was 360, 37 missing)

SELECT count(*) FROM m2m_person_location;
--answer = 2061 (was 2261, 200 missing)

SELECT count(*) FROM m2m_person_occupation;
--answer = 1883 (was 2126, 243 missing)

SELECT count(*) FROM m2m_person_person;
--answer = 2093 (was 2112, 19 missing)

SELECT count(*) FROM m2m_person_religion;
--answer = 593 (no change)

SELECT count(*) FROM m2m_person_society;
--answer = 1238 (was 1387, 149 missing)

SELECT count(*) FROM m2m_person_suffix;
--answer = 1461 (no change)

```

Figure 6.22 Correcting Quaker records

6.19.6 Database views

Views were created to extract data segments from the database to pass manageable sets of data to both JNB and Gephi (see Figure 6.23) and all views built were audited. Audit scripts in SQL were retained and could be referred back to if at any time it was felt that the data had been compromised (see Figures 6.24 and 6.25).

```
DROP VIEW [vw_hddt_ceda_tuples_attributes];

● CREATE VIEW vw_hddt_ceda_tuples_attributes
AS
SELECT (first_names || " " || family_name) AS Name,
       ceda.name AS Target,
       m2m_person_ceda.first_year AS first_year,
       m2m_person_ceda.last_year AS last_year,
       IFNULL(birth_year, 'NA') AS birth_year,
       IFNULL(death_year, 'NA') AS death_year
FROM person
INNER JOIN m2m_person_ceda
        ON m2m_person_ceda.person_id = person.id
LEFT JOIN ceda
        ON ceda.id = m2m_person_ceda.ceda_id
WHERE
        m2m_person_ceda.first_year IS NOT NULL
        AND
        m2m_person_ceda.last_year IS NOT NULL;
SELECT COUNT (*) FROM vw_hddt_ceda_tuples_attributes;
```

Figure 6.23 Data View sample


```

-- There are 5 sets of views:

-- Set VIEW 1 (vw_1_) lists the full data record for persons and also the persons related to each of: club - location - occupation - religion - society.
-- The person views show Names as separate fields (first, last) and also joined (first last in one field.
-- A person table view also shows Quakers marked and another shows Quakers extracted .

-- There are 3094 persons including 592 quakers.

SELECT COUNT(*) FROM vw_1_person_table;
--answer = 3094 - The complete person table
SELECT COUNT(*) FROM vw_1_person_names;
--answer = 3094 - The complete person table with names joined in a single field
SELECT COUNT(*) FROM vw_1_person_with_quakers;
--answer = 3094 - The complete person table with religion joined (marks table with Quakers)
SELECT COUNT(*) FROM vw_1_quakers;
--answer = 592 - The complete person table with religion but only showing Quakers
SELECT COUNT(*) FROM vw_1_religion;
--answer = 592 - The complete religion table
SELECT COUNT(*) FROM vw_1_club;
--answer = 323 - The complete club table
SELECT COUNT(*) FROM vw_1_location;
--answer = 2061 - The complete location table
SELECT COUNT(*) FROM vw_1_occupation;
--answer = 1883 - The complete occupation table
SELECT COUNT(*) FROM vw_1_society;
--answer = 1238 The complete society table

-- Set VIEW 2 (vw_2_) This set shows all memberships of all entities by all persons (9989).

SELECT COUNT(*) FROM vw_2_person_person_relationships;
--answer = 2076
SELECT COUNT(*) FROM vw_2_religion_membership;
--answer = 592
SELECT COUNT(*) FROM vw_2_society_membership;
--answer = 1238
SELECT COUNT(*) FROM vw_2_club_membership;
--answer = 323
SELECT COUNT(*) FROM vw_2_location_membership;
--answer = 2061
SELECT COUNT(*) FROM vw_2_occupation_membership;
--answer = 1883
SELECT COUNT(*) FROM vw_2_ceda_membership;
--answer = 3892
SELECT COUNT(*) FROM vw_2_all_bipartite_memberships;
--answer = 9989
SELECT COUNT(*) FROM vw_2_all_bipartite_memberships_xceda;
--answer = 2007

```

Figure 6.24 Data Views were audited by using the count function (1)

```

-- Set VIEW 3 (vw_3_) This set produces two 'attribute' views where vw_3_all_names_attributes is a SOURCE file for SNA
-- and vw_3_all_bipartite_attributes is a TARGET file for SNA. Both of these files depend on the two 'bipartite' views to compile.
-- The vw_all_names shows the SOURCE (or person) data c/w id.

SELECT COUNT(*) FROM vw_3_bipartite_names;
--answer = 3094
SELECT COUNT(*) FROM vw_3_bipartite_nodes;
--answer = 514
SELECT COUNT(*) FROM vw_3_all_names;
--answer = 3608
SELECT COUNT(*) FROM vw_3_all_names_attributes;
--answer = 3608
SELECT COUNT(*) FROM vw_3_all_bipartite_attributes;
--answer = 9989

-- Set VIEW 4 (vw_4_) Shows only CEDA memberships, and an extract of the Quakers only.

SELECT COUNT(*) FROM vw_4_ceda_membership_dates;
--answer = 3892
SELECT COUNT(*) FROM vw_4_ceda_membership_quakers;
--answer = 643

-- Set VIEW 5 (vw_5_) shows the person to person relationships.

SELECT COUNT(*) FROM vw_5_person1;
--answer = 2080
SELECT COUNT(*) FROM vw_5_person2;
--answer = 2080
SELECT COUNT(*) FROM vw_5_person1_person2;
--answer = 2080

```

Figure 6.25 Data Views were audited by using the count function (2)

6.19.7 Visualising data using JNB and Gephi

The P7 Report comprises several JNB files rendered in PDF format and then combined into one report. This is a static report with no functionality and significantly reduced data visualisation clarity. The best way to see the HDDT is in operation through its component parts, DBeaver, JNB and Gephi, and the best rendition of a P7 project report is as an HTML file rendered in a browser. The P7 Report has been uploaded to its own dedicated website: <https://kelvinbeerjones.github.io/project-seven-book/intro.html>. It will later be offered to the Quaker Studies Research Group as its first wholly digital study.

Data can be visualised in JNB by building charts using the Python Pandas library.⁵⁴⁵ The P7 Report uses set styles and consistent choice options throughout to ease in reading and comparison of one chart to another, and one visualisation to another. For example, in the P7 Report only bar and pie charts are used, and they are displayed consistently. Data relationships are visualised in Gephi, which has impressive flexibility options for data visualisation (see Figure 6.26).

⁵⁴⁵ https://pandas.pydata.org/docs/user_guide/visualization.html (Accessed 23/2/2035)

Actions / step	1	2	3	4	5	6
Layout	Force Atlas					
Appearance	Nodes	Colour	Ranking	Degree		
Statistics	Network Diameter					
Appearance	Nodes	Size	Ranking	Betweenness Centrality	Min 1	Max 35
Layout	Adjust by sizes	Repulsion Strength				
Statistics	Modularity					
Appearance	Nodes	Colour	Partition	Modularity Class		
A	Node Size	[Slider]				
Filter	Topology	Degree Range	[slider]			
Preview	Show labels	[adjust size] = 24				
Preview	Export	png				
Save	Project					
Image size	1280w					

Figure 6.26 Gephi options set-up

The GexF file in Gephi was used to generate a graph file using the Force Atlas algorithm, the 'network diameter', 'modularity' and 'appearance' routines. This topology arrangement was found, by a process of exploring a range of possible topologies, to produce a suitable network graph for analysis. Gephi's modularity routine clearly displays large society group networks and its betweenness centrality routine reveals both individuals and the smaller satellite groups which play important roles linking groups together. It should be noted that the visualisation displays in this thesis result from establishing a Gephi topography and are the result of running the above algorithms; the graphs have not been manually arranged on the page.

Finally, the graphs were saved as 'project files' allocating the Network display to the JNB container reserved for each exercise. PNG image files of the graphs and selected areas of the graph were generated to produce the P7 Report and thesis figures.

6.19.8 Visualising relationships

The composition of the CEDA data and the non-familial relationships between CEDA members can be represented and studied visually (see Figure 6.27). It can be seen that this is a highly dense network of bipartite relationships. (Quaker-to-Quaker relationships are omitted from this graph.)

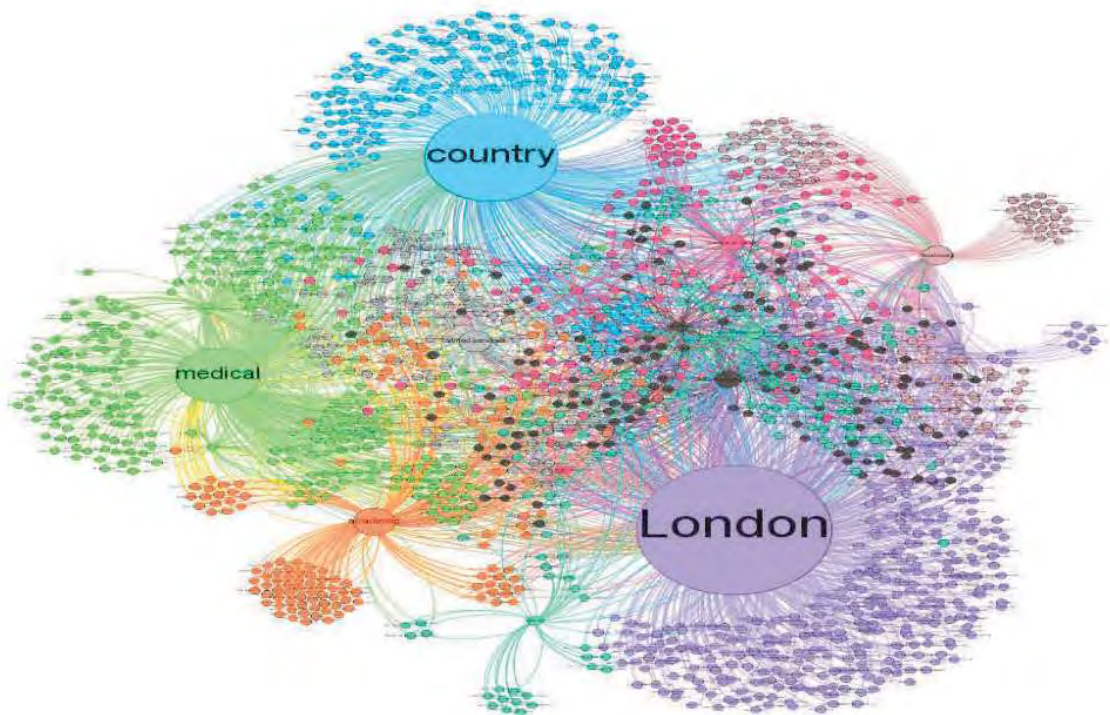


Figure 6.27 The relationships in the CEDA database

Two locations are most prominent in the graph, London and Country, and they are analysed further here. In the RAI data 'Country' refers to 'Britain excluding London'. London and Scotland are other UK locations, and these three locations can be combined in an analysis to emphasise the importance of Britain as a whole in the CEDA. The prominence of Britain in

the data is not surprising given that the CEDA represents those whose object was to support the discipline of anthropology in Britain.

Relatively smaller (but more interesting) clusters are only just visible: Medical, Armed-Services, Athenaeum Club and Geological Society. To better see the CEDA community, visualisation graphs of smaller segments of the data were created.

6.19.9 Visualising the CEDA members

Figure 6.28 shows the members of the CEDA, beginning with a small group of people in the QCA (bottom left). Most of the QCA members soon join the APS, which was created shortly after the QCA. Two members of the QCA bypass the APS and join the ESL. The ESL (bottom right) appears next, with its origins clearly coming from the APS. Two clusters emerge on the network stream from the APS to the ESL and these represent those APS members who formed the ESL. Then many APS and ESL members later formed the AI. However, before that occurred the APS members who joined the ASL can be seen. The members of the ESL and the ASL finally merged to form the AI, which became the RAI, in 1871. The clusters of key individuals who bridge these networks, all of which originate in the APS, are a clear indication that (1) the anthropological societies in the CEDA are highly networked and (2) members of the non-scientific group, the APS, played a significant role in the development of the CEDA.

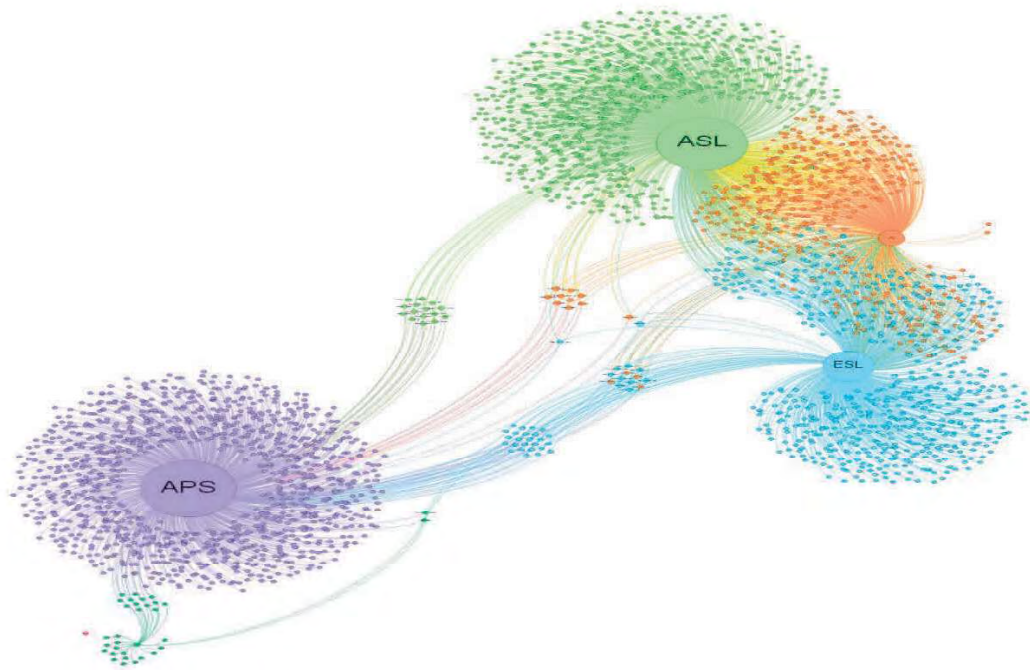


Figure 6.28 Members of the CEDA

6.19.10 Locations

A close look at locations brings to light the extent of networking (and support) the CEDA had from comparable societies in other countries. This is borne out by a reading of the Notes sections of the RAI data for the persons that the HDDT has identified by Gephi based on data originating from an SQL select query. The RAI Notes frequently name the foreign societies each of these people represent. The non-British members of the CEDA are shown both in a JNB pie chart (see Figure 6.29) and a Gephi graph (see Figure 6.30). This combination of charts and graphs can be used to analyse the same data and is an example of the complementary analytics that the HDDT affords. The data in both the JNB chart and the Gephi graph are the same, but they reveal different information. The data in both can be analysed in full detail in the SQL database using DBeaver select queries.

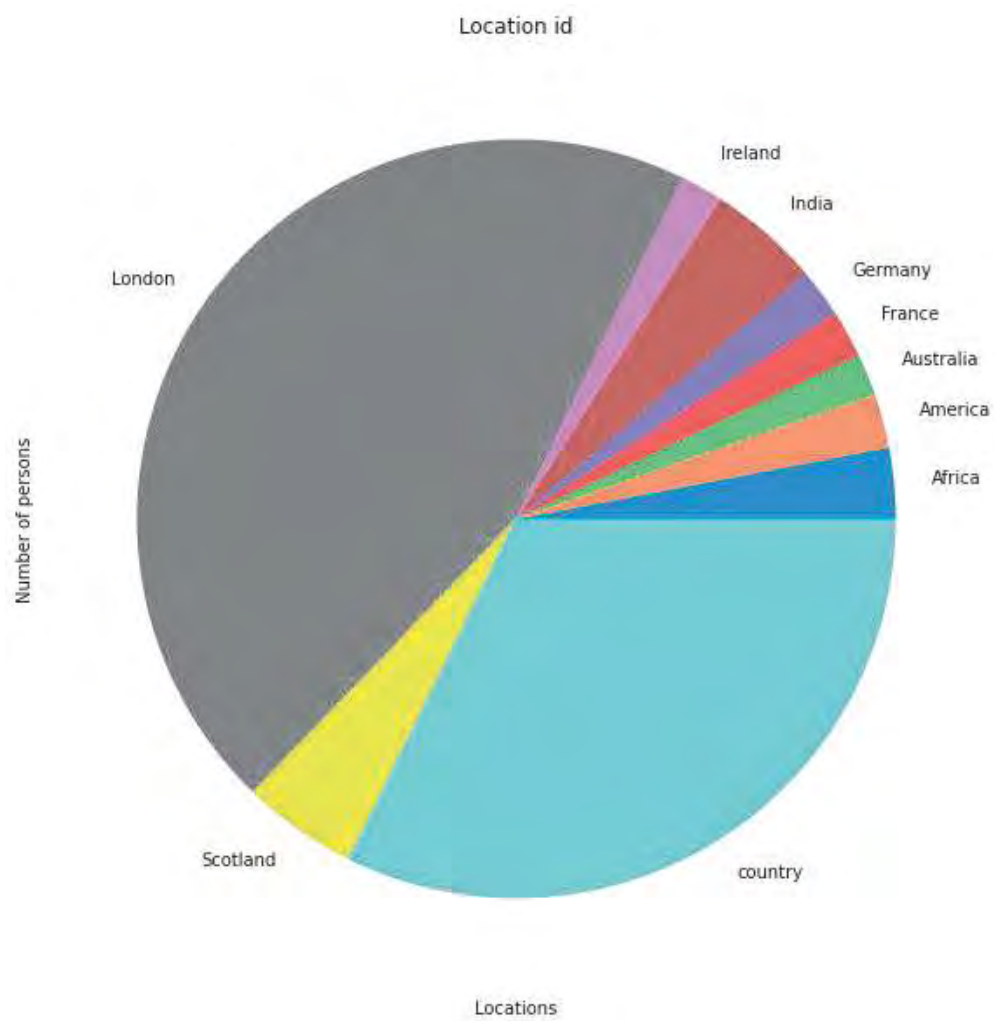


Figure 6.29 Locations pie chart

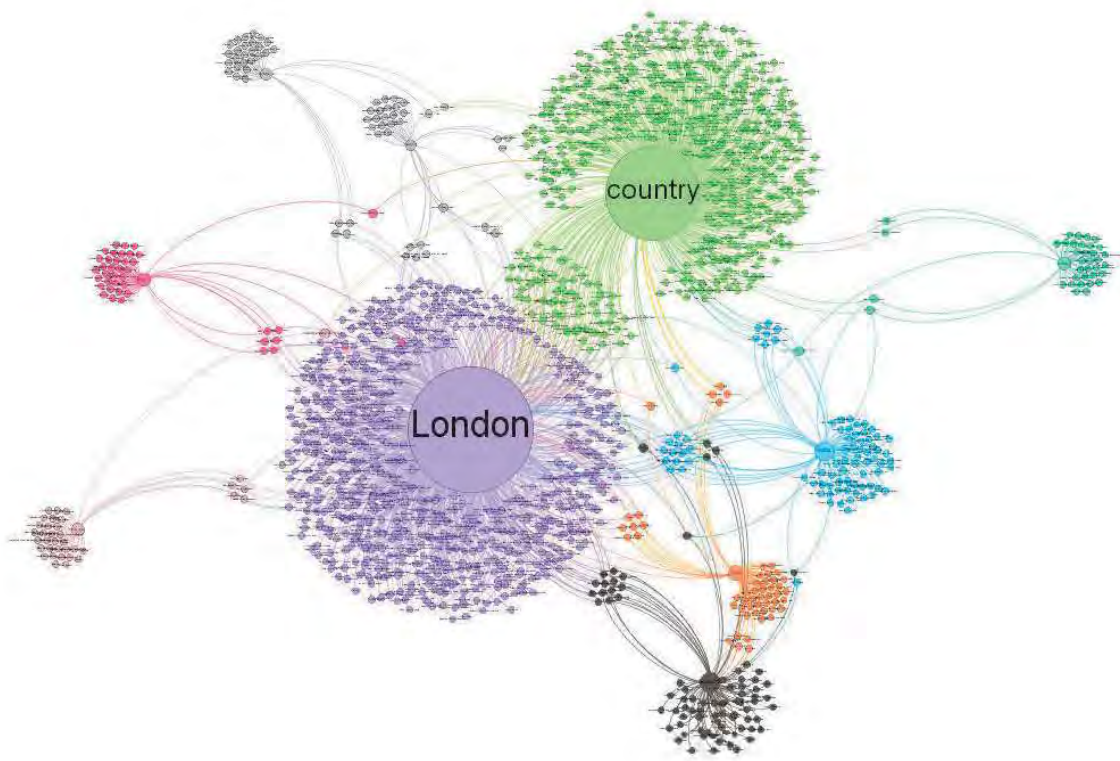


Figure 6.30 Networks span many nations

6.20 Case Study 1: The Centres for the Emergence of the Discipline of Anthropology (CEDA)

6.20.1 Introduction: the work of Ronald Rainger

Scholars have generally recognised the nineteenth century as the period in which the various fields of scientific enquiry became separate and well organised disciplines. In England particularly, an abundance of specialised societies, institutes, and associations had sprung up by mid-century, robbing the Royal Society of its former dominance and making London a virtual capital of science. Among the disciplines organised at this time was anthropology, the scientific study of Man. In the years 1837–1871 anthropology emerged, both institutionally and intellectually, from its early association with religious and social concerns to become a natural science among the natural sciences. (Rainger 1976, Abstract)⁵⁴⁶

This study looks back to the work of Ronald Rainger in 1975 and 1980 on the emergence of the discipline of anthropology in Britain, and enhances it with a digital study of the networks between the CEDA members.⁵⁴⁷ Rainger is absent from the post-1990 published literature but his work is apposite to this study. He also did not have the benefit of the Hodgkin Family papers, which were not catalogued by the Wellcome Institute until after 1990.⁵⁴⁸ Rainger

⁵⁴⁶ 'Two features characterise and qualify this study. First, the emphasis of this work is on the organisational development of anthropology. Obviously, any scientific society is organised around an intellectual tradition, and the first chapter is an attempt to indicate the problems and questions which defined anthropology and influenced its later organisation. Nevertheless, I have concentrated on the institutional, not the intellectual history of this discipline. Though these two dimensions of the science cannot be completely divorced, certain important aspects of nineteenth century anthropology had little effect on the institutionalisation of that science and thus have but a minor place in this study. On the other hand, because non-scientific factors have also defined anthropology throughout much of the nineteenth century, the differing attitudes of scientists, philanthropists and racists towards non-European peoples in an era of imperialism directly influenced that institutionalisation. This work is an examination of that influence, a study of the societies created by these groups of men and the part each played in the organisational development of anthropology' (Rainger 1976, Abstract).

⁵⁴⁷ The term CEDA is the author's, not Rainger's.

⁵⁴⁸ 'The majority of the Hodgkin papers are in the hands of Mrs. Dorothy Hodgkin, London, and Professors Thomas and Dorothy Hodgkin, Oxford, England. Both Professor Kass of Harvard University and Dr Ruth Hodgkinson of the Wellcome Institute for the History of Medicine, London, possess copies of some eleven

recognises the establishment of the ESL in 1843 as the beginnings of a scientific approach to anthropology,⁵⁴⁹ but he also recognises that the roots reach further back into the early 1830s.⁵⁵⁰

Rainger saw Hodgkin as a key organiser in the disciplinisation of anthropology in Britain, but argues that Hodgkin's interest began with his evidence at the Select Committee on the Aborigines in 1832, whereas this thesis shows Hodgkin working at that time within Quaker committees (the QCA) that parallel the APS.

Rainger picks up traces of Hodgkin's activity in the QCA and the APS only in Hodgkin's correspondence.⁵⁵¹ He recognised Hodgkin's dual interest in faith-based philanthropy and science and how they both significantly pre-dated Hodgkin's participation in the work of the Select Committee,⁵⁵² and also how intimately Hodgkin's twin interests appeared to be

spools of microfilm of those materials. In conversation with Dr Hodgkinson on May 11th, 1973 however, I was informed that both the microfilm edition and the Hodgkin materials themselves are still in an unclassified state, and therefore an examination of them would be a time consuming and possibly fruitless enterprise. For that reason, my research has not included analysis of the correspondence or private papers of Dr Hodgkin' (Rainger 1976, 36).

⁵⁴⁹ 'The establishment of the Ethnological Society of London in February 1843 marks the beginning of the organised scientific study of man in England. Scientific enquiry in ethnology, as in the other descriptive, natural historical sciences, was not intellectually isolated from the religious and social milieu; that study, as defined in the work of James Cowles Pritchard and expanded upon by the Ethnological Society of London, at least prior to 1859, entailed certain theological and social assumptions which were reflected not only in that work, but in the make-up of the organisation as well' (Rainger 1976, 63).

⁵⁵⁰ 'The British and Foreign Aborigines' Protection Society, an organisation which arose through the efforts of individuals who are not significant figures in the history of anthropology and whose interests were related but largely peripheral to that science. Their concern was philanthropic, directed to the welfare of non-European peoples. Nevertheless, the Aborigines' Protection Society was the first organisation in England whose concerns even touched on anthropology, and an analysis of that society offers some insight into the nature of anthropology and the nature of organised institutions in early nineteenth-century England' (Rainger 1980, 792).

⁵⁵¹ 'By virtue of an extensive correspondence maintained with Friends in various British colonial settlements, Hodgkin was well aware of the cultural contact situation throughout much of the world. Information from these sources convinced him that the advance of European civilisation was having a most deleterious effect on non-European peoples, in certain cases threatening to exterminate them entirely (Hodgkin 1830–1866; Society of Friends 1838–1842' (Rainger 1980, 705).

⁵⁵² 'The Aborigines' Protection Society was founded in 1837, but the incentive to establish such an organisation had already existed for some years. This was first made manifest by Dr Thomas Hodgkin (1798–1866), the eventual founder, whose works and ideas best summarise the motives and objectives of the organisation.

connected.⁵⁵³ This study shows that Hodgkin acted on those interests a decade before the formation of the ESL. However, in recognising that Quakers played a formative role in the APS, Rainger does not make the networking linkages that lead the QCA to stand out as the origin of the CEDA.

Quaker involvement moved from the QCA to the APS at its formation in 1837, with the QCA being laid down five years later in 1842. Rainger notes that the Quaker influence in the business of the APS caused Hodgkin to noticeably subordinate his scientific interests to his Quaker philanthropic concerns.⁵⁵⁴ In 1843, a year after the QCA was laid down, Hodgkin and a group of other APS members resolved this conflict of interests by forming the ESL, which was firmly a science-based society and had no philanthropic agendas, although that society did not easily shake off the philanthropic concerns of its founders.⁵⁵⁵

Hodgkin, for many years a demonstrator in morbid anatomy at Guy's Hospital, London, was a man with widely divergent interests and concerns. Primarily these included matters of a medical, political or religious nature. In 1835, however, he delivered a paper to the Philological Society of London on the importance of the analysis and preservation of languages, particularly non-European languages. This interest in non-European peoples had scientific as well as humanitarian roots' (Rainger 1980, 703).

⁵⁵³ 'The ways in which philanthropic and scientific objectives coexisted and competed in the work of this organisation indicate that neither the intellectual nor the institutional boundaries between science and nonscience were as well defined in early nineteenth-century England as they are today' (Rainger 1980, 703).

⁵⁵⁴ 'A statement put forward by Sir Culling Eardley Smith and agreed to by the Society recognised the importance of ethnological investigation, but still considered that scientific activity as the "most availing means of arresting the progress of evil, and aiding in the introduction of the blessings of Christianity" (APS1842: 6)' (Rainger 1980, 710). 'The collection of ethnographic materials [by the APS] was initially a part of the Society's active attempt to reorient British colonialism' (Rainger 1980, 709).

⁵⁵⁵ '[M]en with a more extensive background in science such as Ernst Dieffenbach and Richard King were not enthusiastic about the continued priority given to humanitarianism, especially in the light of the political failures and financial setbacks of the Society. The lip service accorded ethnology did not placate those who felt the need to cultivate science for its own sake and on its own terms. So King, a physician and a naturalist, who at the time was secretary of the organisation, issued a prospectus for an ethnological society to the Aborigines' Protection Society on July 20, 1842. Excluding reference to any colonial political objectives, he outlined proposals for the institutionalising of what he termed the study of the natural history of man: the collection and classification of materials on the races of man, the collection of the most authoritative works on ethnology for reference and the use of people travelling abroad, the distribution of funds to such travellers, and the establishment of a correspondence with similar societies and with persons residing in remote, predominantly non-European countries (King 1844: 15–16). King's proposal was not accepted by the Society; consequently, he, Hodgkin and the few others interested in the cultivation of ethnology began work on their own to organise a separate institution. They in fact comprised only Hodgkin, King and William Aldam, each of who maintained a dual affiliation with both societies' (Rainger 1980, 711).

6.20.2 A static view of the CEDA network

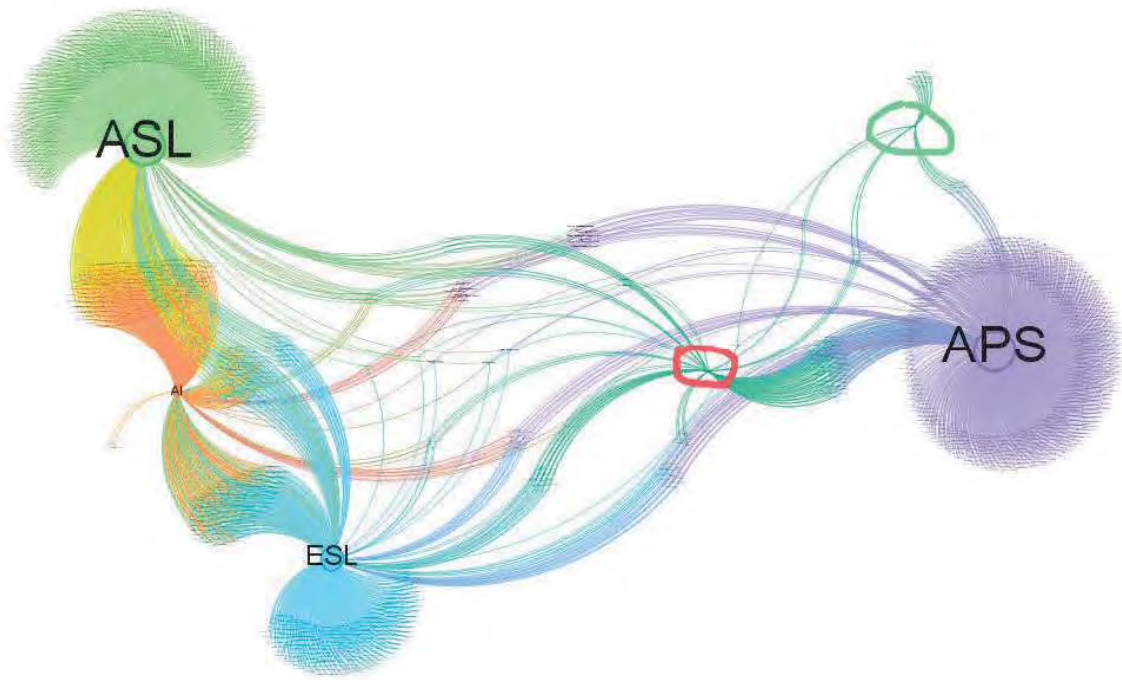


Figure 6.31 The CEDA network with the QCA marked in green and Hodgkin's network marked in red

Continuing from the introduction to the CEDA network in Section 6.19.10 where locations were analysed, this section focuses on the CEDA networks themselves. Figure 6.31 shows the CEDA network. There is evidently a high level of shared memberships in and across the CEDA community. The QCA is highlighted in green and Thomas Hodgkin's personal mixed Quaker/non-Quaker network is shown highlighted in red. (This personal network will be examined in detail in Case Study 3, see Section 6.22.) The flow through time in the graph is generally from right to left. This Case Study shows that the QCA, under the leadership of

Hodgkin, is the initiating organisation of the CEDA because it is the prime mover in the establishment of the APS. At its establishment, a large number of Quakers join the APS and then later a few Quaker members of the APS help to form the ESL and the ASL, but now working within intentionally secular societies. There is considerable dual membership between the ESL and the ASL, and when these two societies merge in 1868 to form the AI, shortly after Hodgkin's death, the influence of the QCA is at an end and the Quaker interest in the APS also begins to decline.

6.20.3 A JNB analysis of the QCA

The QCA began as a Quaker Committee to consider the State of the Heathen in 1832⁵⁵⁶ and then restructured in 1837 to become the Committee on the Aborigines. There were ten active members, rising to twenty-four in 1837.⁵⁵⁷ (Appendix 3.3 discusses in detail the shifting objectives of the QCA.) Reports on the activities of the QCA exist only in occasional mentions in the Quaker minutes of the annual Meeting for Sufferings. Minutes or reports of the QCA itself cannot be located (see Figure 6.32). This Friends House Quaker archivist's note records that 'No manuscript minutes or reports of the Aborigines Committee have been traced (Oct 1976)'.

⁵⁵⁶ 'Concerned to take the Gospel to the heathen.' This study does not consider this earlier committee, it instead focuses on the 1837 revision to the committee's objectives to the plight of the aborigines.

⁵⁵⁷ There were sixty-four members at the committee's formation in 1862, but this study has only considered the few members active in subsequent years.



Figure 6.32 Archivist's note on the QCA Friends House Archives

The most important entry in the record of the Quaker Meeting for Sufferings comes in June 1839, when we learn that Thomas Hodgkin has joined the Quaker Committee on the Aborigines: 'The list of correspondents has now been called over, and the following alterations and additions have been proposed and agreed to viz; "In London, Thomas Hodgkin added."' ⁵⁵⁸ The next day Hodgkin formed the secular APS at Ratcliffe Quaker Meeting House (see Appendix 1).

Figure 6.33 illustrates the durability of the QCA with only one or two new members each year adding to the initial group, and Figure 6.34 shows that eleven of the committee

⁵⁵⁸ Meeting for Sufferings, 7th of the 6th month 1839, pp. 587 and 588.

members left after the committee was restructured in 1839, and none after that, until the committee was laid down in 1847.

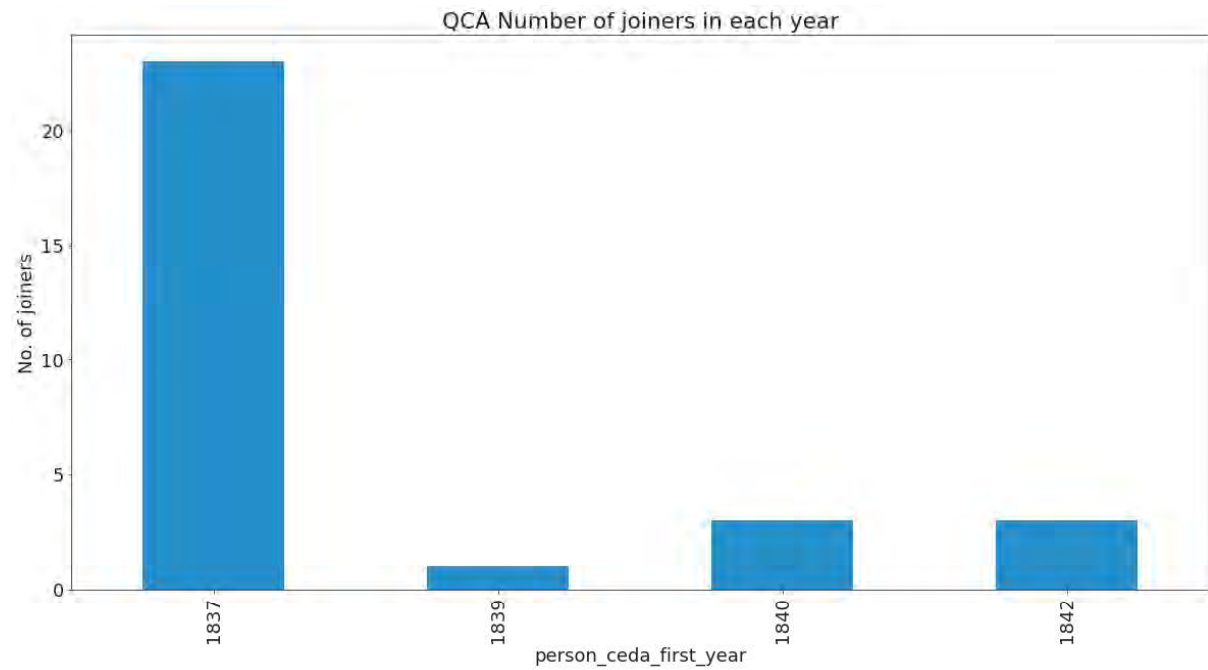


Figure 6.33 QCA joiners by year

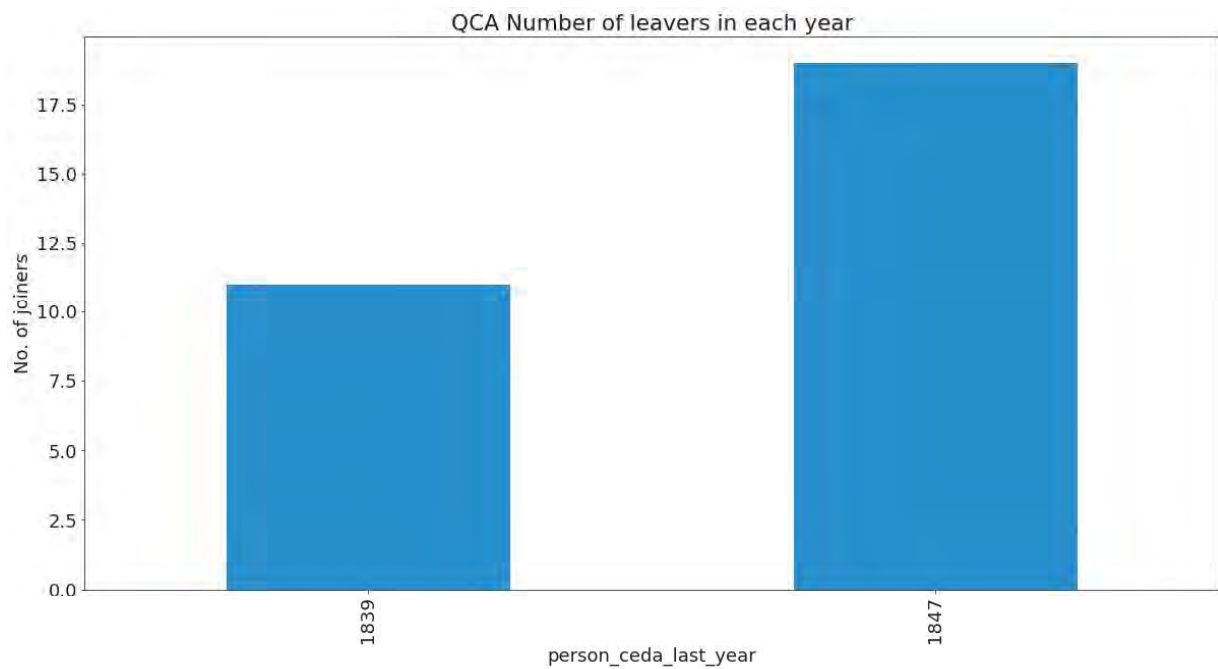


Figure 6.34 QCA leavers by year

6.20.4 A JNB analysis of the APS

Figures 6.35 and 6.36 show APS leavers and joiners in each year. The high take-up of memberships in 1838 and 1839 (315) is partially offset by the 120 leavers in 1840. (Data for the years 1841–1846 are irretrievably missing.) When the data picks up again in 1847, there seems to be a period of stability in memberships from then until 1851, when 115 leavers are replaced by a similar number of joiners. Then from 1860 onwards a period of change arises with large numbers of both joiners and leavers in each year.

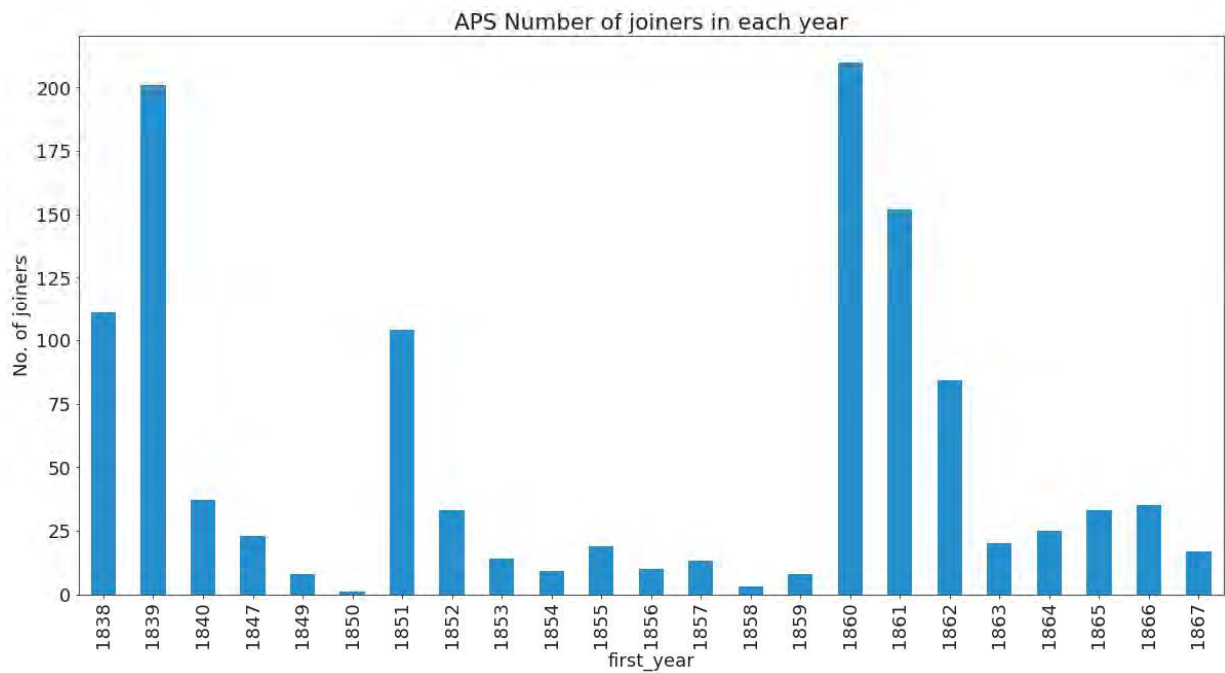


Figure 6.35 APS joiners in each year

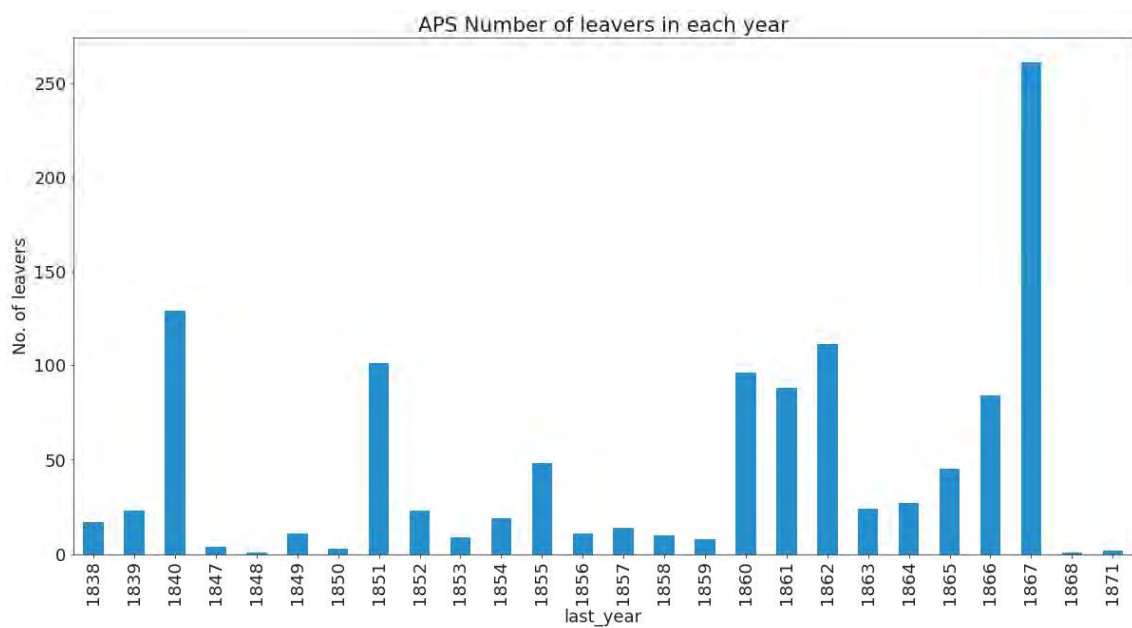


Figure 6.36 APS leavers in each year

Greater insight into these uncertain years, 1840, 1851 and the 1860s, resides in the Quaker data for the same period. Figure 6.37 and 6.38 show Quaker joiners and leavers of the APS.

Quakers join the society in large numbers at its formation in 1838 and 1839 and only in 1840 is there a Quaker turnover, resulting in a net 12 leavers. The turnover in Quaker joiners and leavers in subsequent years is low until 1860, with the exception of 1851, when over 50 Quakers join the APS. In 1860, about 110 Quakers join, with high joiner numbers in the following two years too. In those years the number of leavers also rises, but proportionately with the number of joiners. In 1867, the year after the death of Thomas Hodgkin, 190 Quakers leave the APS.

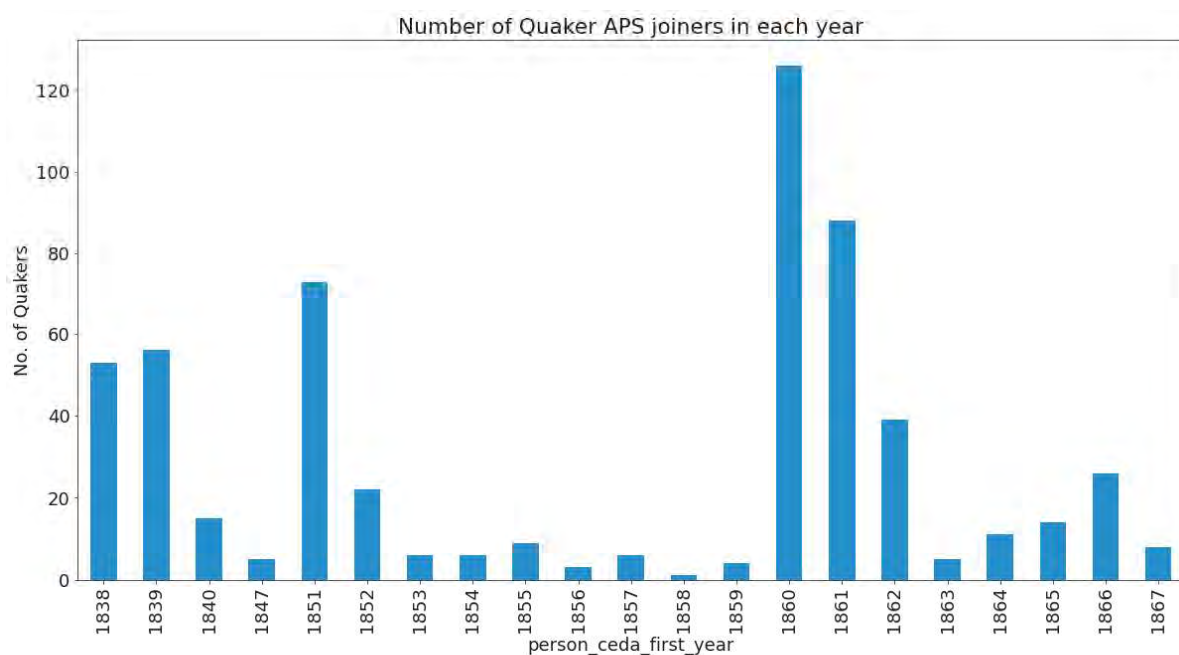


Figure 6.37 APS Quaker joiners in each year

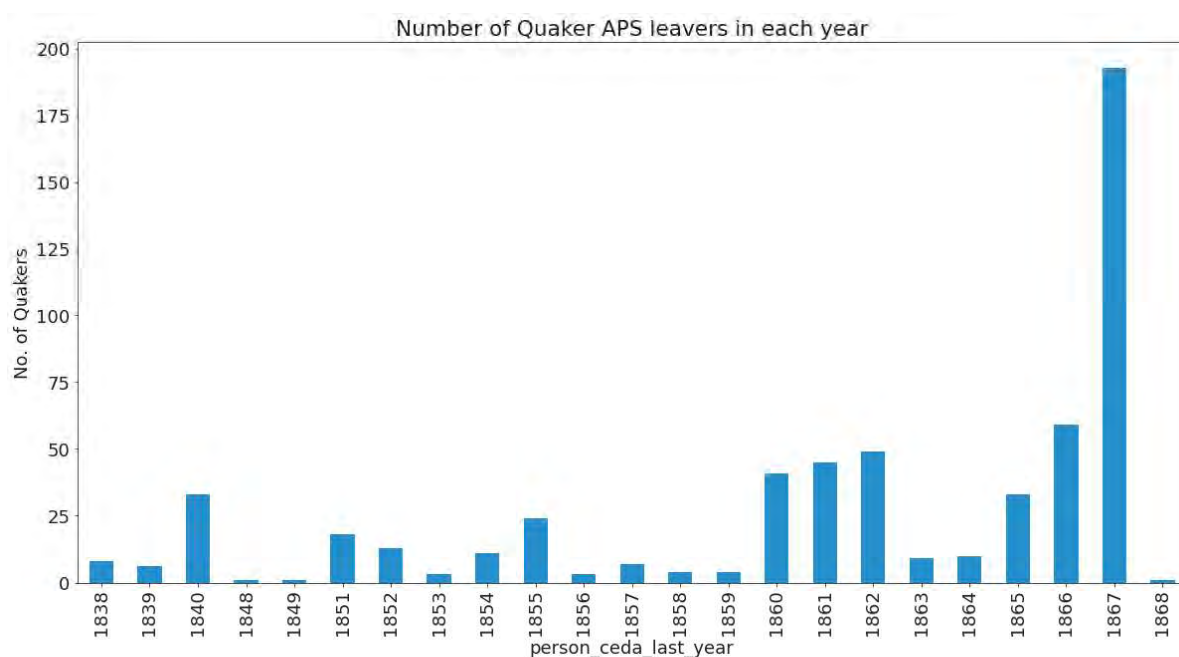


Figure 6.38 APS Quaker leavers in each year

There seems to be a correlation between the membership turnover in the APS in general and the turnover among Quaker members. In 1838 half of the joiners are Quaker and in 1839 only a third. This would indicate that while the Quakers played a large role in the first formative year of the APS, many non-Quakers joined soon after. The high number of leavers in 1840 are not predominantly Quakers, but some Quakers do also leave the APS that year. In 1851 most of the large number of joiners are Quaker. In 1860 half of the large number of joiners are Quaker and in 1867 nearly all of the leavers are Quaker.

Mindful that only 20% of the CEDA are Quakers, the data in these charts indicates that Quakers are exerting considerable influence in the business of the APS at critical moments. More investigation is needed into the activities of the APS in 1850/51 and 1859/60 to discover what this data indicates on the ground, but in the literature there is nothing of note that might explain the membership movements of 1851. However, it is possible that in

1861/62 the APS was preparing to birth a new society, the Anthropological Society of London, for example to ensure votes favourable to the transaction, which was effected in 1863 .

6.20.5 A JNB analysis of the ESL and ASL

The number of Quakers in the ESL is only five in 1844, and one or two in each year thereafter. There are twice that number of Quakers in the ASL in all years, but still a relatively low number. This suggests that Thomas Hodgkin's network has a large Quaker presence in the APS and a medical/scientific presence in the ESL and ASL. Thomas Hodgkin's networking in support of both his philanthropic and scientific interests is taken up in Case Study 3 (see Section 6.22).

6.20.6 Case Study 1 conclusion

Question 1 revisited

Can the model reveal the networks between the members in the five organisations that comprise the CEDA? This question is important because it resolves a wider and current uncertainty over the origins of the discipline of anthropology in Britain and the extent of Quaker involvement.

The HDDT was primarily designed to answer this question and it has done so. The CEDA networks have been shown, analysed and visualised. The Quaker presence in the network and how that presence influenced the make-up of the APS at critical moments have been

observed. The linkages between the QCA and the APS have also been shown, reinforcing the argument that the origins of the CEDA lie in the QCA and in particular in the role of Thomas Hodgkin MD. The study was successful, even though a considerable amount of data is missing. After extensive enquiries it must be assumed that this data is lost. But it could materialise, in which case consideration of that data might add more detail and even reveal further interesting movements of people in and out of the CEDA, although it is unlikely to significantly change the nature of this study.

6.21 Case Study 2: The CEDA Quakers and their relationships

This Case Study should be read in conjunction with Case Study 1, which illustrates the Quaker memberships of the CEDA. The HDDT holds records for 589 Quakers and among them there are 2006 family relationships: 1265 distant relationships, 500 close relationships and 241 intimate relationships.⁵⁵⁹ Functional limitations in relational databases limit working with genealogical data. For instance, in genealogical software hierarchical relationships are seamlessly merged with lateral relationships, enabling a large range of possible relationships to be modelled, such as second cousin twice removed. SQLite is unsuitable for this type of relationship modelling.⁵⁶⁰ The same constraints apply to JNB and Gephi. This is the weakest part of the HDDT and shows the difficulty of assembling one suite of technologies to tackle a wide range of data types. What follows is a necessary simplification of the available family relationship data and with limited functionality due to the limitations of the HDDT. Nonetheless, it is possible to view the family relationships among the 589 Quakers in the CEDA at sufficient granularity for this Case Study.

⁵⁵⁹ Distant relationships are coded '1' in the HDDT to indicate a family relationship further than '2' or '3'; close relationships (coded '2') are cousins and immediate relationships (coded '1') are for parent/child/sibling relationships. Quakers can have relationships in all three categories.

⁵⁶⁰ This is one of the main reasons why genealogists use paid-for software. FamilySearch, WikiTree and Gramps are perhaps the only popular free software, and each of these uses complex software to build family relationships.

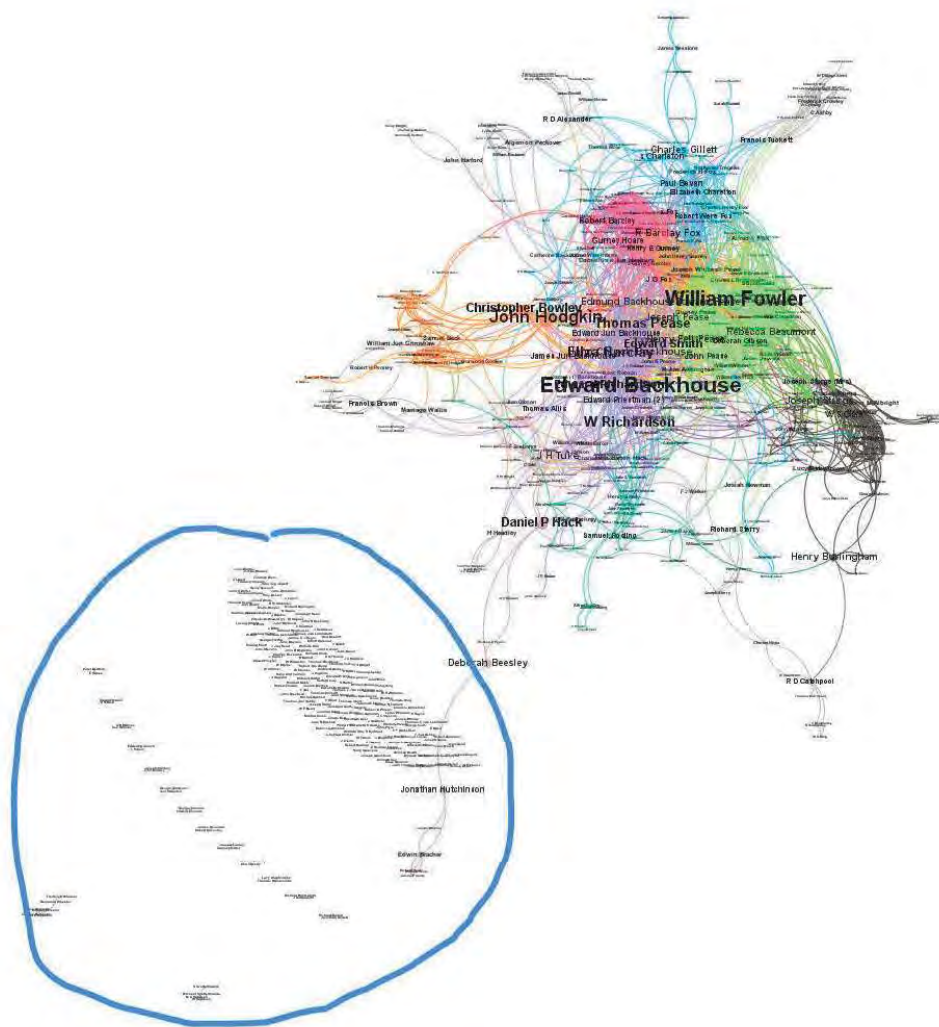


Figure 6.39 All Quakers and their family relationships – circled in blue are no or only one relationship

Figure 6.39 shows all of the Quakers in the HDDT with their family relationships. Circled in blue are those approximately 100 Quakers who have no family members present, those who are related to only one other member, and those with very small, isolated family networks among the CEDA membership. For example, Jonathan Hutchinson has a small network of his own which is displayed separately to the intensely networked members. He is related to Deborah Beesley and she is better connected to the larger Quaker community. Therefore, Jonathan Hutchinson's small network is shown connected to the larger networks through his relationship with Deborah Beesley. It is significant that 17% of the Quaker

members of the CEDA are members through either interest or friendship rather than family. Friendship is a category of particular interest when studying Quakers (otherwise known as the Society of Friends). It may well be that there are other close relationships among the Quakers that have no family with them – they may instead have ‘Friends’ here who are fellow members of Quaker Meetings or Quaker Committees. P7 did not seek out this data and so it is beyond the scope of P7 to explore this further.

Figure 6.40 shows that network graphs in Gephi can be filtered to reduce the number of plots, which enables particularly dense parts of the Quaker family networks to be better displayed.



Figure 6.40 Filtering graphs in Gephi

6.21.1 Quaker relationships by type

The chart of Quaker relationships by type (Figure 6.41) shows that Quaker family relationships by relationship type are roughly equal between immediate and close relationships and these two taken together are larger than those Quakers with distant relationships.

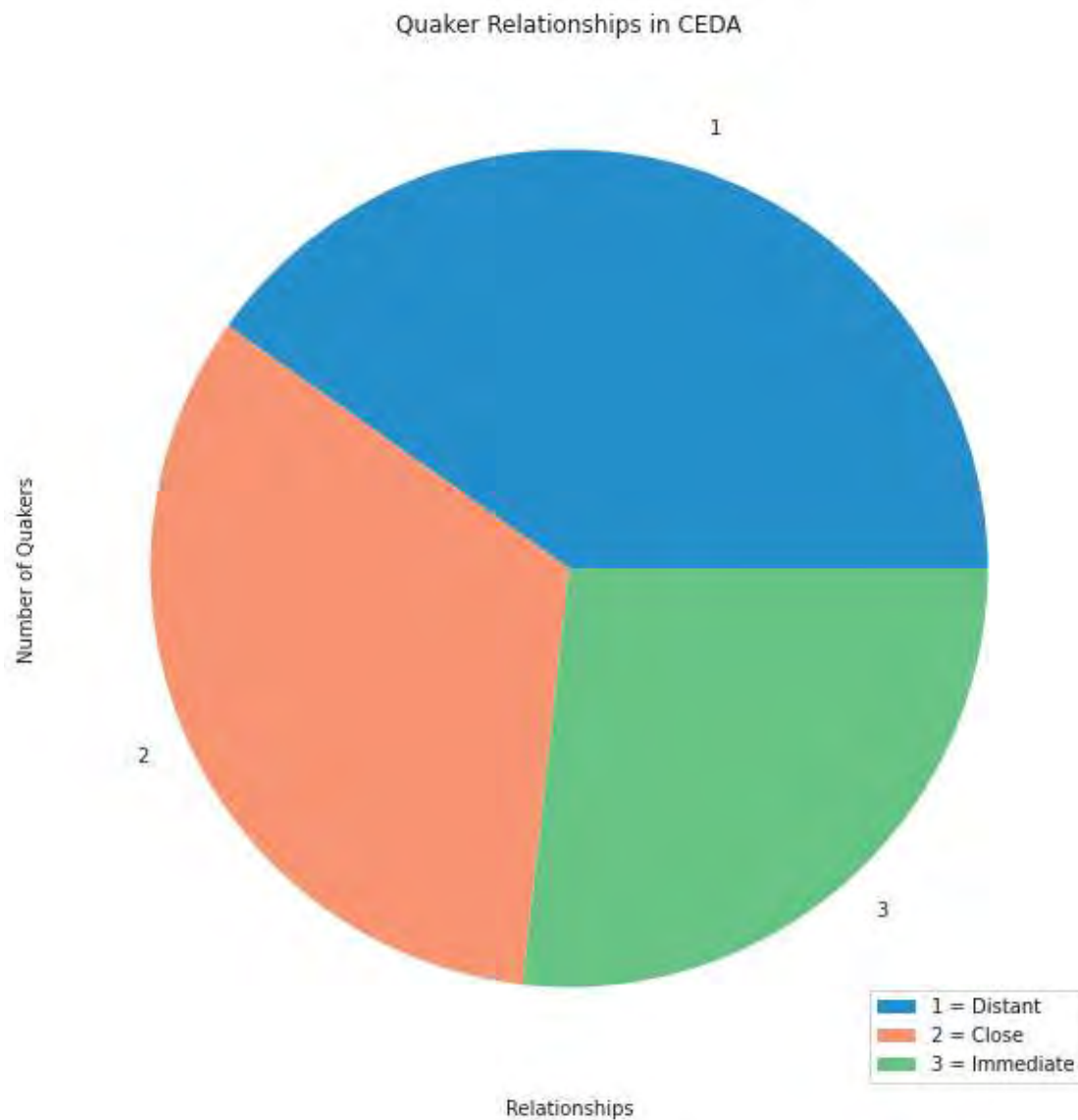


Figure 6.41 Quaker relationships by type

6.21.2 Quaker CEDA memberships

Figure 6.42 shows that Quakers in the CEDA are mostly to be found in the APS. This Case Study therefore focuses on the family relationships between the Quaker members of the APS.

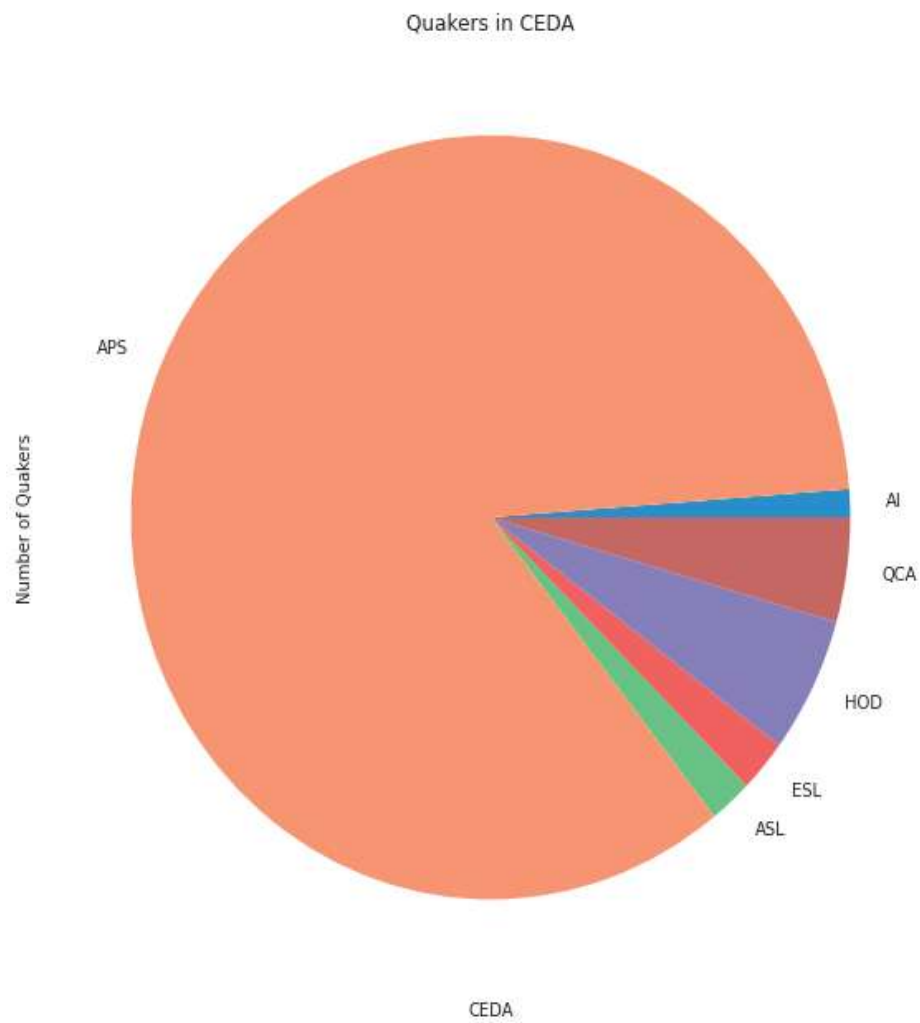


Figure 6.42 Quaker presence in the CEDA

6.21.3 Quaker members of the APS – distant relationships

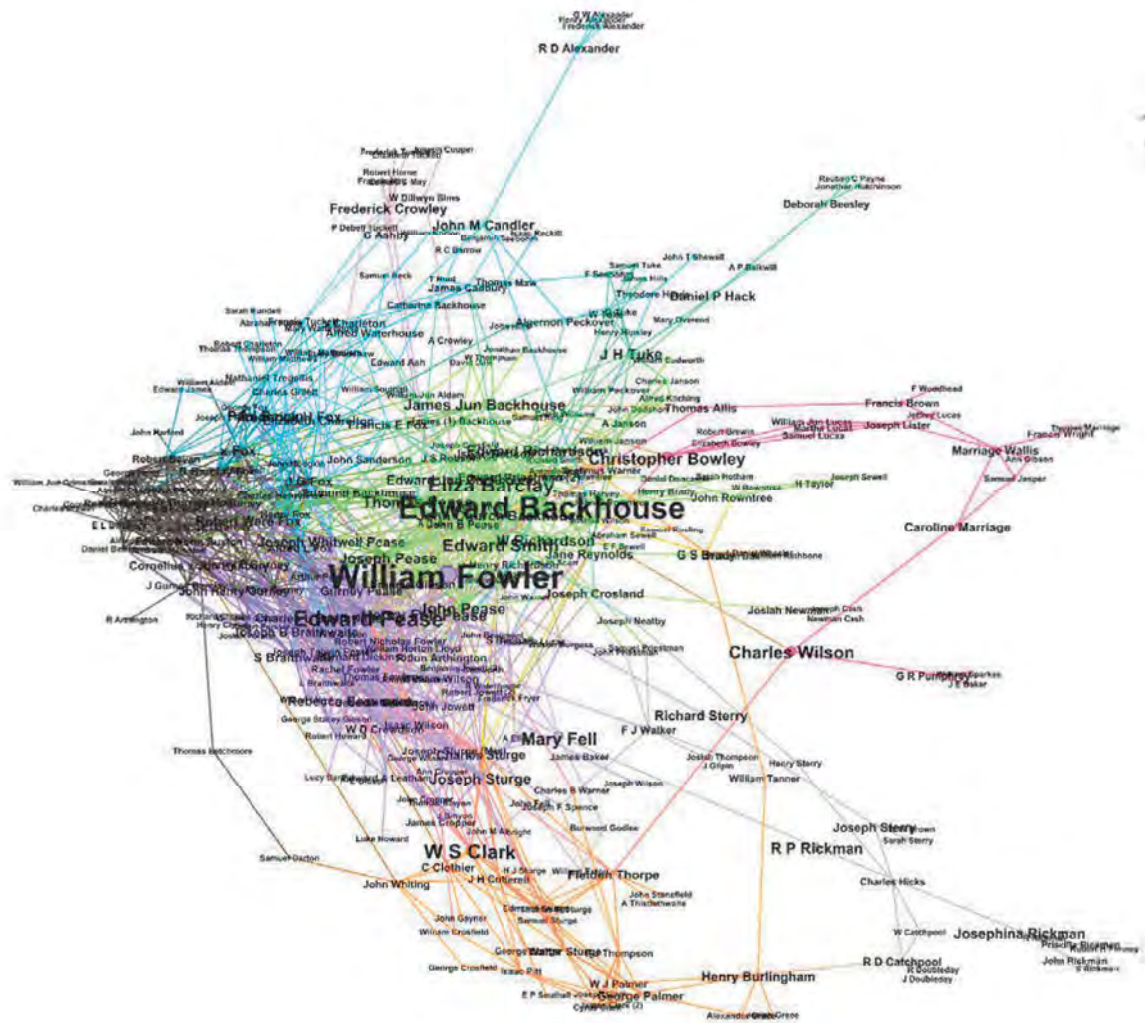


Figure 6.43 Distant relationships

Figure 6.43 illustrates the complex interconnectivity between Quakers with distant relationships. These form into large groups of Quakers with connected sets of distant relationships, as revealed by the colour coding in the graph.

6.21.4 Quaker members of the APS – close relationships

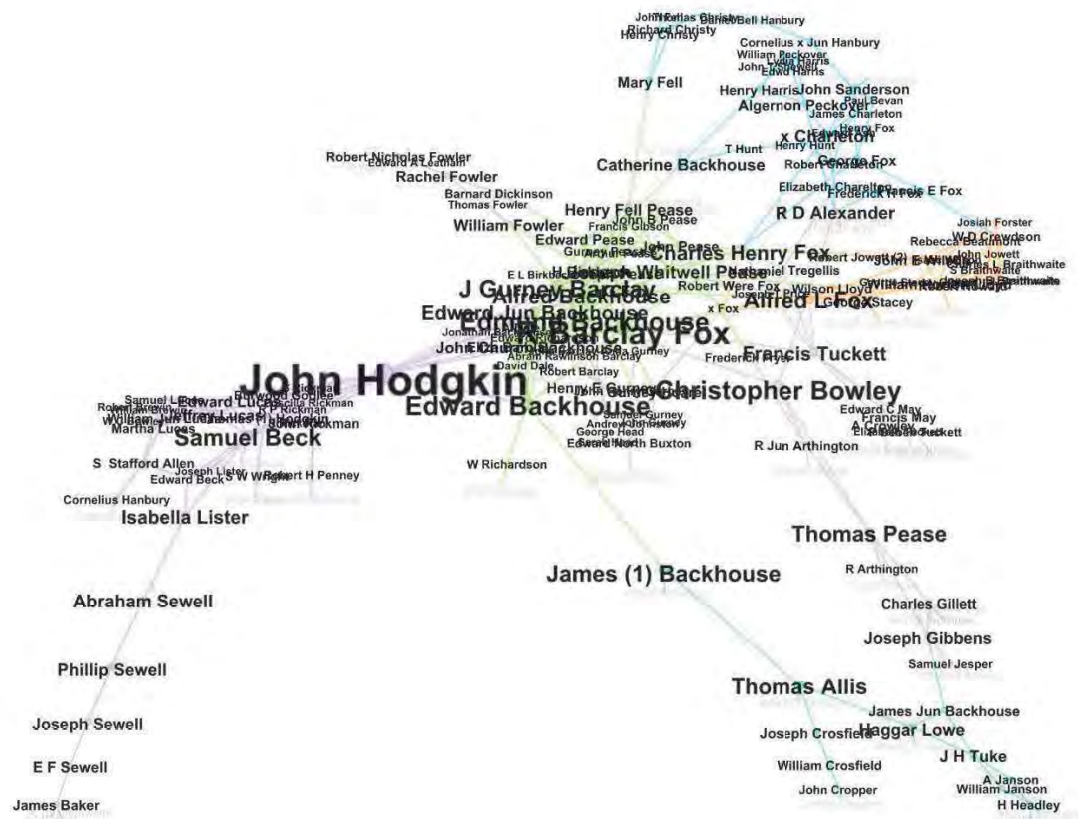


Figure 6.44 Close relationships

Figure 6.44 shows close relationships among the Quakers. John Hodgkin, Thomas Hodgkin's nephew, a close family member and fellow journeyman in the CEDA, has the most close family connections. Notice how four members of the Sewell family string out at the bottom left of the graph. They are connected to each other and then through Abraham Sewell to Isabelle Lister, who then has extensive family connections with the main network centred around R. Barclay Fox.

6.21.5 Quaker members of the APS – immediate relationships

Figure 6.45 shows Quakers with immediate family relationships among the CEDA membership. This is perhaps the most remarkable graph because it shows a central mass of persons, most of whom have only one close relationships in the community. Appearing like satellites around the large central cluster are those with ever greater numbers of close family members. R. Barclay Fox's, Edward Backhouse's and John Hodgkin's families are nested together to the left of the chart. They are nested because they have also have family connections between them. Then rotating clockwise round the graph are the Rickmans, the Crosfields, the Gibson/Beaumonts, the Sturges, the Sessions, the Gurneys, the Charletons and the Peases. Many of these are prominent Quaker families socially and politically active in several areas at this time.

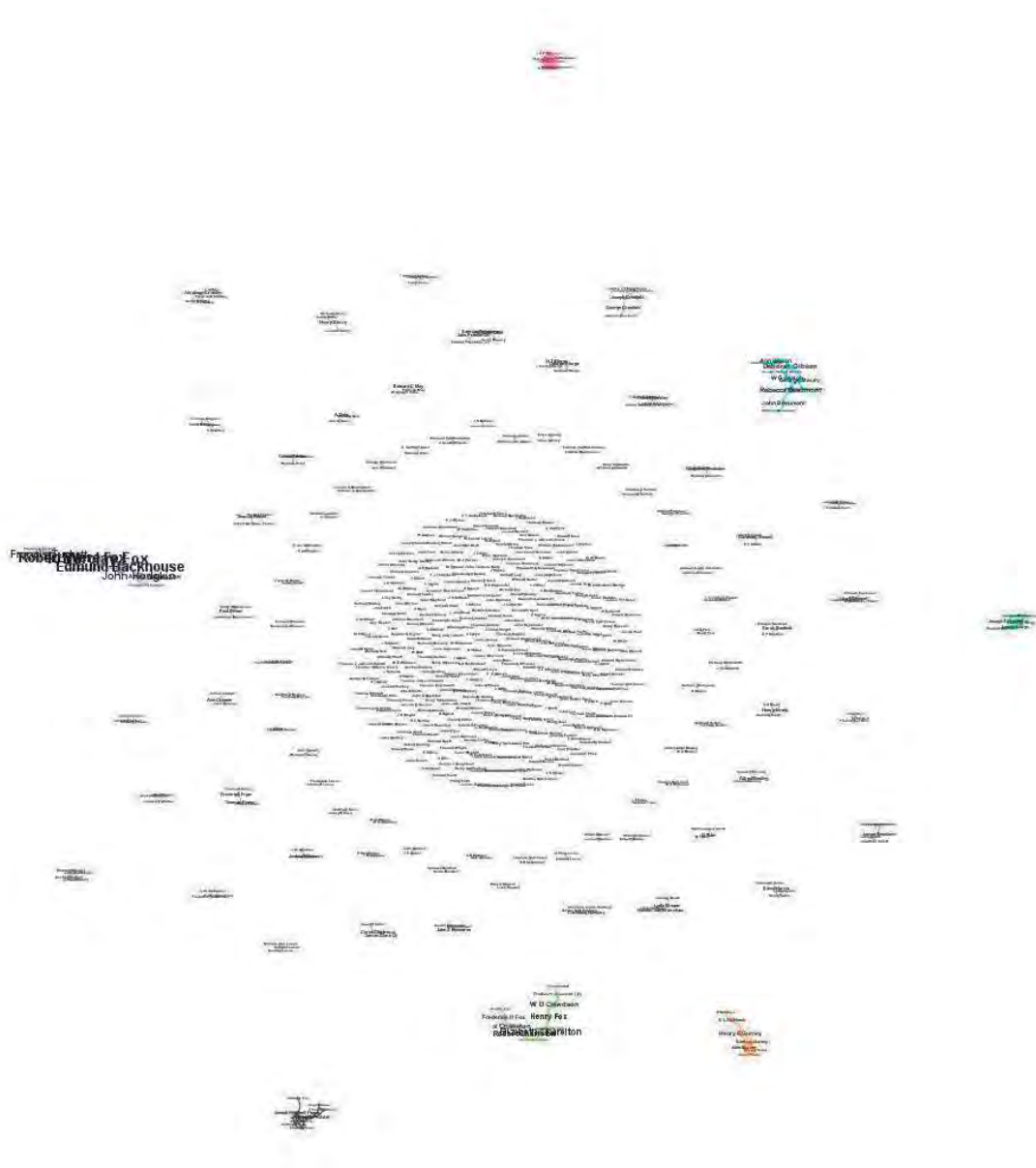


Figure 6.45 Immediate relationships

6.21.6 A selection of individual networks

Figures 6.46–6.49 show the family relationships of a selection of those prominent in the network graphs above. Immediate relationships appear in green, close relationships in red and distant relationships in mauve. Thomas Hodgkin has only one close relationship, with his nephew John. Thomas married late in life and had no children; John Hodgkin married several times and had at least twelve children. Edward Backhouse had several close and

many distant relationships, bringing him into familial contact widely across the CEDA.

William Fowler had the largest number of close relationships.

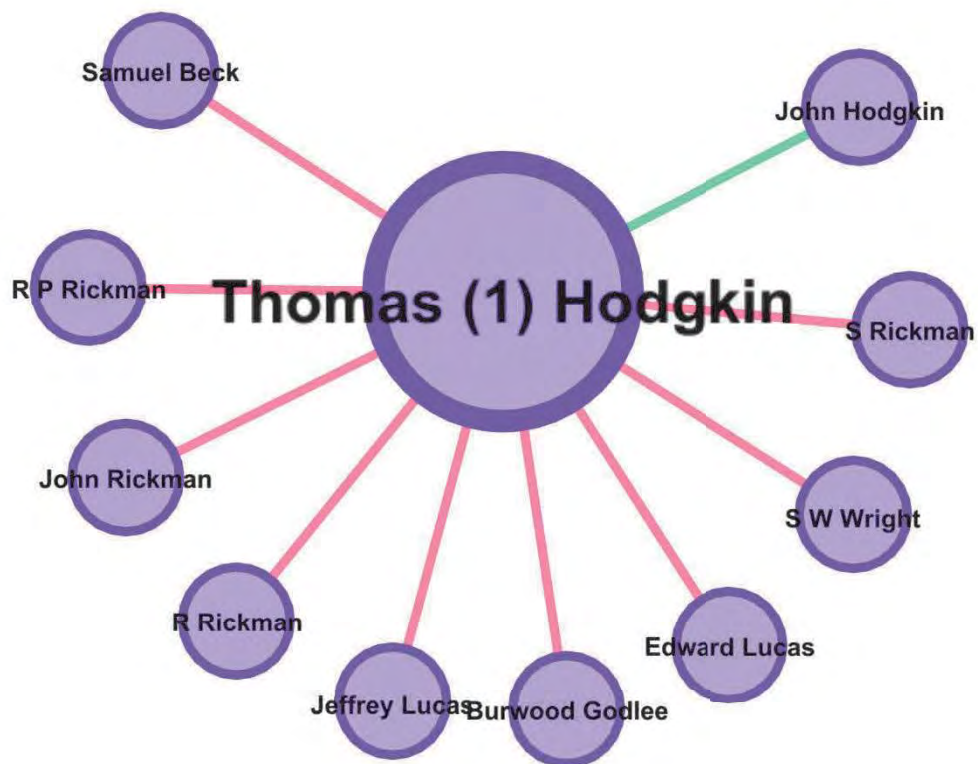


Figure 6.46 Thomas Hodgkin's family relationships

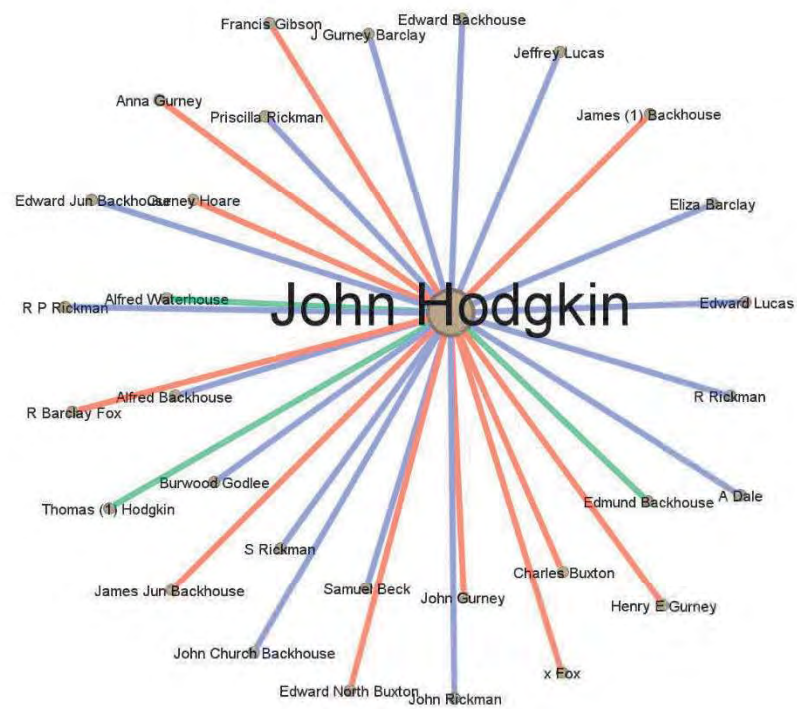


Figure 6.47 John Hodgkin's family relationships

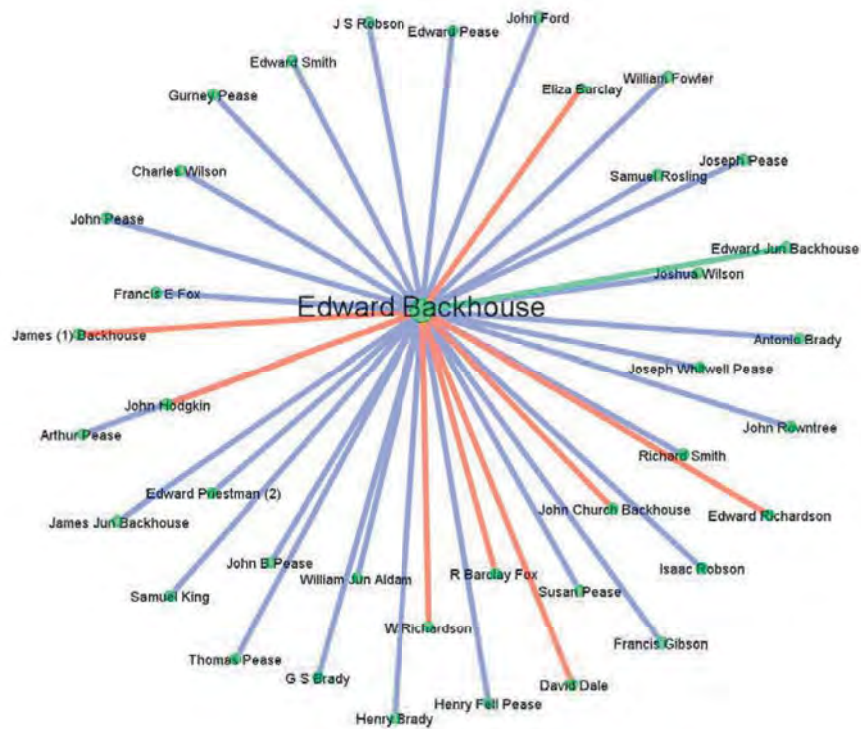


Figure 6.48 Edward Backhouse's family relationships

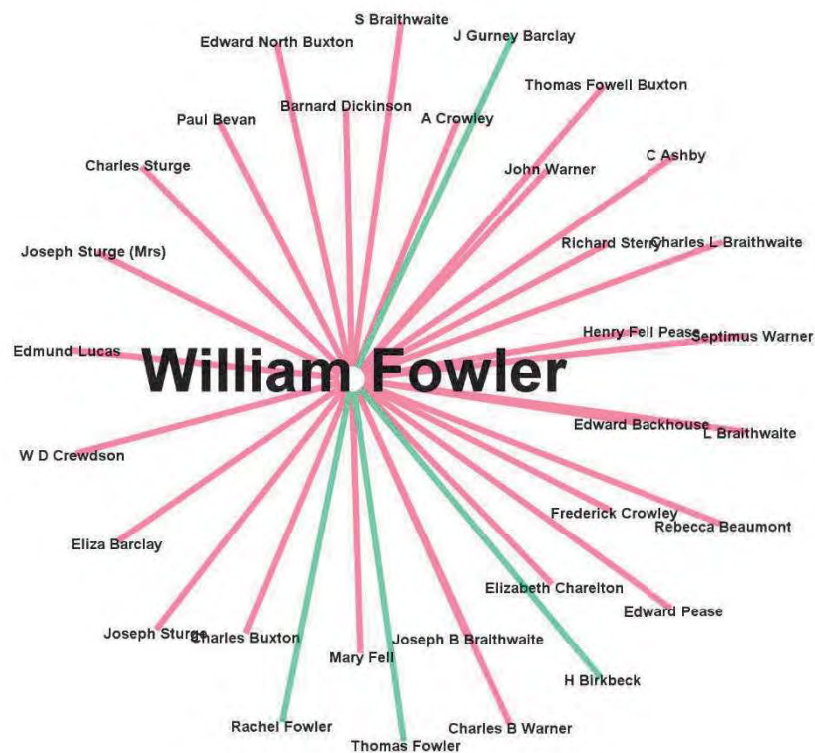


Figure 6.49 William Fowler's family relationships

6.21.7 Case Study 2 Conclusion

Question 2 revisited

Can the model examine Quaker-to-Quaker relationships and how these relationships supported the Quaker members of the CEDA during the forty years of its life?

The HDDT was not designed to handle genealogical data; the decision to consider a family relationships exercise emerged as a result of undertaking the other two exercises.

Consequently the exercise had to be limited in scope, because neither SQLite, JNB nor Gephi could work with Gedcom files. This is a lesson in project creep. Nonetheless, by simplifying

the family relationships to immediate, close and distant family members, much meaningful detail has emerged that supports Case Studies 1 and 3. It is in this supporting role that Case Study 2 is best considered. The study has identified key families and persons who have significant presence in the network graphs and are worthy of further study. The answer to Question 2 is yes – but only partially.

6.22 Case Study 3: Thomas Hodgkin's MD networks

This chapter examines one of the major social networks embedded in the CEDA, that of Thomas Hodgkin MD 1799–1866. The index to Laidlaw's *Protecting the Empire's Humanity: Thomas Hodgkin and British Colonial Activism 1830–1870* (Laidlaw (2021)) was extracted to identify persons appearing in both the index and the CEDA database. The hypothesis is that if someone is a CEDA member and also appears in the index, then it is reasonable to consider them as members of Hodgkin's network.⁵⁶¹ In this Case Study the *Protecting the Empire's Humanity* network is called PEH.

Laidlaw draws extensively on the Hodgkin Family Papers Collection in the archives of the Wellcome Institute, London. The catalogue to this collection includes an index of correspondence sent by Hodgkin and an index of correspondence received by Hodgkin. The two indexes were extracted and combined to find persons who exchanged correspondence with Hodgkin (correspondences both received and sent). The hypothesis was that if there was an exchange of correspondence, then the correspondents, if they were also present in

⁵⁶¹ That the term of this study and Laidlaw's publication cover the same research area time-frame is because this study was born out of discussions with Laidlaw (see Chapter 1).

the CEDA database, could be considered as members of Hodgkin's network.⁵⁶² In this Case Study the Wellcome Institute collection network is called WEL. The Case Study explores the centrality of Hodgkin's personal network within the wider CEDA network. Will a close examination indicate who the key influencers in Hodgkin's network might be?

6.22.1 Introduction to Case Study 3

The 'Introduction' section in (Laidlaw 2021) sets out the importance of the social networks of Thomas Hodgkin MD: 'The roots of this book lie in the personal correspondence of the Quaker, scientist, and activist, Dr Thomas Hodgkin' and 'The exploration of Britain's imperial history presented in this book is profoundly shaped by Thomas Hodgkin's personal papers, today housed in the Wellcome Library in London, and what they reveal of his philanthropic, medical, and scientific interests and networks. The volume and the breadth of that archive has allowed me to trace and assess a wide array of influences on 'imperial humanitarianism' (2021, 3). Laidlaw also says that Hodgkin's personal archive at the Wellcome Institute constitutes the 'backbone' of the book and that 'Its study reveals 50 years' worth of unlikely connections, sustained relationships, and closely argued, if sometimes contradictory, cases, (2021, 7).

⁵⁶² This Case Study excludes one-way correspondence between any two persons, such as where the recipient of correspondence never replies to it and/or never writes back directly or even indirectly to the sender. An example might be correspondence from/to a CEDA member which could be seen as intensive lobbying of a representative of the Colonial Office, simply because he was perhaps considered to be 'in charge', or had published on matters of interest to Hodgkin and/or his networkers, or who was simply thought to be susceptible to influence, but where the recipient shows no interest in networking directly with the sender.

Laidlaw does not, however, rely solely on the Wellcome Institute collection, and in the 'Acknowledgements' section she details other primary archives used in the research.⁵⁶³

Laidlaw says she has also 'drawn heavily' on what she calls 'the most important' biography of Hodgkin, that by Amalie M. Kass and Edward Kass (Kass 1988),⁵⁶⁴ as well as the papers of the APS held in the Bodleian archives.⁵⁶⁵

Only describing Hodgkin's networks as text in narratives makes it difficult to 'see' the people who formed the extensive networks in which Hodgkin operated, how many people there were and how they related to each other.⁵⁶⁶ And so, *Protecting the Empire's Humanity*, while discussing at length the efforts of Hodgkin's networks to relieve the plight of aborigines throughout the British Empire, does not make those networks visible or make

⁵⁶³ 'Much of the early archival work for this book took place in the archives of the Wellcome collection in London. I am grateful to the knowledgeable and helpful reading room staff and archivists both there and at other libraries and archives on which I depended: the Baillieu library, University of Melbourne; the Bedford library, Royal Holloway; the British Library; the National Archives at Kew; the library of the Religious Society of Friends, London; Rhodes House library, University of Oxford; the Royal Anthropological Institute; the Royal Geographical Society; and the State Library of Victoria' (Laidlaw 2021, xi).

⁵⁶⁴ 'Hodgkin's life is not a new subject for historians, and I have drawn heavily, and with gratitude, on the most important biography of Hodgkin, "Perfecting the World", by the historian of medicine, Amalie M Cass and her medical doctor partner, Edward Cass. Though written over 30 years ago, perfecting the world was also grounded in Hodgkin's archive (then in private hands)' (Laidlaw 2021, 3).

⁵⁶⁵ 'Hodgkin was, and until his death in 1866 would remain, the central figure within the Aborigines Protection Society: to a significant degree, its campaigns were shaped by the priorities he identified, drew on his correspondence networks, and were often supported by his financial largesse ... Most of the organisational records, whether minutes of committee meetings, account books, or registers of correspondence, that documented the early decades of the Aborigines Protection Society no longer exist. Their remnants – mainly correspondence with informants – are housed within the Antislavery Society collection at the Bodleian library in Oxford. However, together with the society's extensive publications, and, above all, Thomas Hodgkin's exceptionally rich personal archive, those remnants connecting the stuttering, but highly revealing, Aborigines Protection Society to the domains of empire and imperial humanitarianism, as well as medicine, science, religion, trade, and politics. Alongside the society's scant early records in Oxford its memorials and addresses to imperial and colonial governments remain in national collections, and were almost always published by the society. In general, the society's publications were numerous: they included both annual reports and accounts of annual general meetings, and a range of stand-alone pamphlets. From 1847, the society published a periodical, "The Colonial Intelligencer or Aborigines Friend". This appeared at intervals varying from monthly to (in 1863–4) biennial; its name also fluctuated: for consistency it is referred to as The Colonial Intelligencer throughout this book. Before 1847, the society had episodically published "Extracts from the papers and proceedings of the Aborigines Protection Society" (hereafter extracts) and sought to place accounts of its work and findings in different newspapers and periodicals, as is explored in chapter 2. Taken together, these sources hint at networks that included colonizers and colonized in every inhabited continent' (Laidlaw 2021, 6).

⁵⁶⁶ See Appendix 9.2.

them easily available to scrutiny. This Case Study fully addresses those two needs and presents new insights about Hodgkin’s networks that would be difficult to achieve without an HDDT.

6.22.2 The PEH network

All of the persons named in the PEH index were extracted from it (Laidlaw 2021, 359). Of the 290 persons named in the index, 108 were found to be already present in the CEDA database. The remaining 182 do not appear in the database and so were ignored. Some of those indexed in PEH will not be members of Hodgkin’s support network, but rather persons that Laidlaw has referenced in PEH for other reasons (for example, King William IV, Queen Victoria).

All persons PEH	PEH persons CEDA	PEH persons not CEDA
290	108	182

Table 6.2 Comparison of persons PEH in Case Study 3

Note: At least 15 persons among the 182 non-CEDA persons appearing in the index to PEH could nonetheless be family members of persons who are CEDA members (from the frequency of shared names). Nonetheless, the intention here is to discover how the 108 who are recorded in the HDDT database relate both to each other and to the wider group of 3000 persons already in the HDDT.

6.22.3 The WEL network

In the Wellcome Hodgkin Family Archive Data Indexes, there are 107 person Hodgkin writes to who also write to him, and of these 46 appear in the HDDT database and 61 do not.

Following the rationale described above for using the PEH index, we can also reasonably assume here that some of the persons in the WEL index will be persons related to members of persons appearing in the HDDT database. Many will not – for example, we can expect to find much medical-related correspondence in an archive collected by the Wellcome Institute (which is a medical history archive) – but these are outside the scope of this thesis, which only analyses political activism among the members of the CEDA.

All persons WEL	WEL persons CEDA	WEL persons not CEDA
107	46	61

Table 6.3 Comparison of persons WEL in Case Study 3

The Case Study 3 data from both WEL and PEH are shown in Figure 6.50.

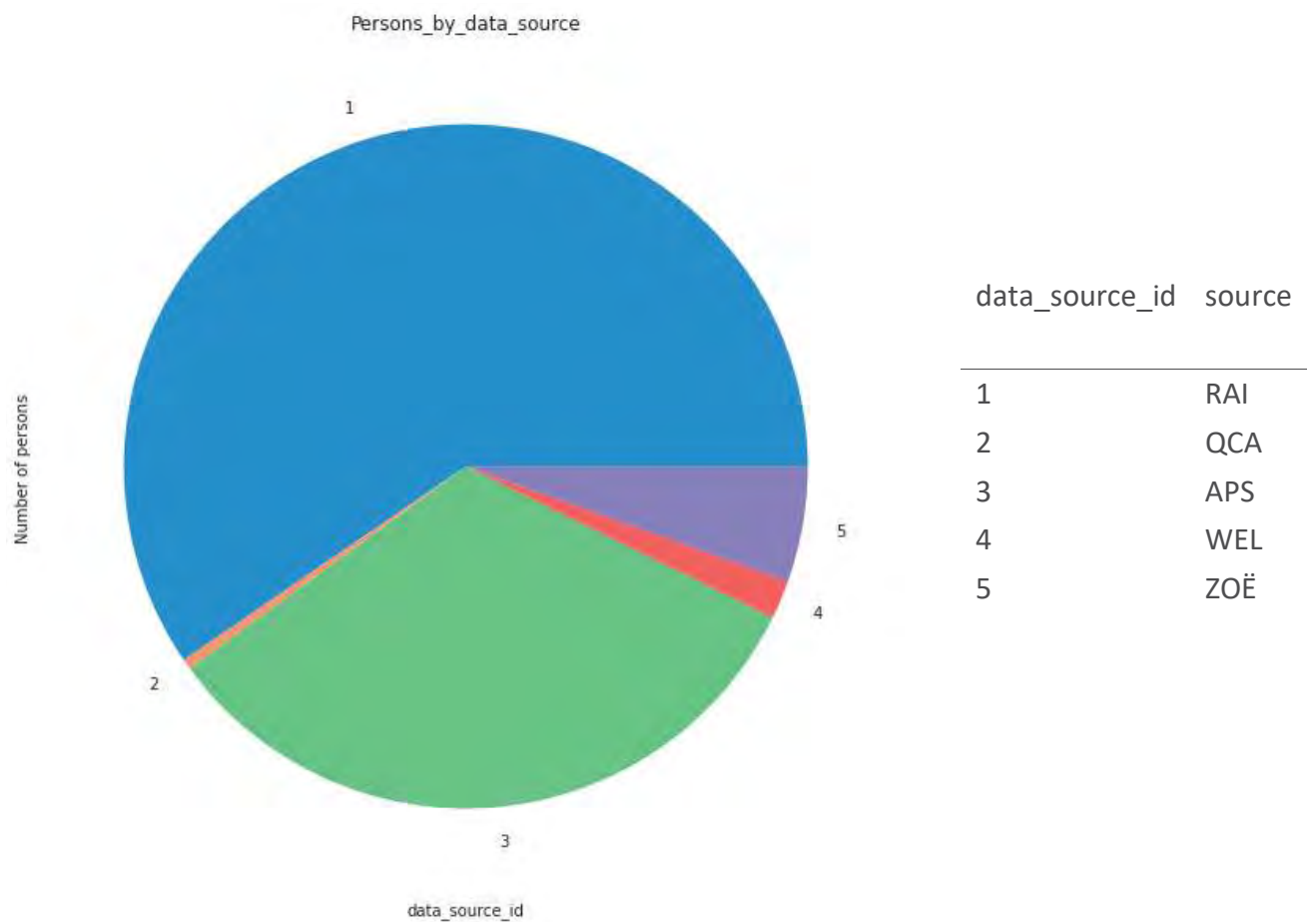


Table 6.4 Case Study 3 data sources

6.22.4 Adding PEH and WEL to the database

The Index to PEH lists 108 persons who already present in the HDDT CEDA database. They were allocated to a 'dummy' CEDA group (table CEDA, Target = 'ZOË'). This enables the visualisation of Laidlaw's Hodgkin political activism network and shows its relationships to

the original HDDT CEDA groups. Using the platform DBeaver (<https://dbeaver.io>), the HDDT CEDA database was modified to accommodate the exercise:

1. Add to person_data_source table two new temporary data sources – ZOË and WEL.
2. Add to CEDA table two new CEDA groups – ZOË and WEL.
3. Person_table. Upload 182 new records from a CSV file of Laidlaw references and allocated them to data_source = ZOË.
4. Person_table. Upload 61 new records from a CSV file of Laidlaw references and allocated them to data_source = WEL.
5. Update m2m_person_ceda table to allocate 108 persons to a new CEDA group = ZOË.
6. Update m2m_person_ceda table to allocate 46 persons to a new CEDA group = WEL.

Once the data from both ZOË and WEL had been added to the HDDT database, the network analysis could be performed and a report drawn up in the form of a Jupyter Notebook.

6.22.5 The CEDA social network including ZOË and WEL

The two ‘dummy’ groups (ZOË and WEL) appear in the network graph (see Figure 6.51) in orange. Also in orange is the QCA group. The QCA is displayed close to the new dummy groups because Hodgkin began his political work in the QCA and all three groups are Hodgkin groups. In the graph ZOË is shown in the centre, WEL to the right and QCA to the left.

Society	Abbrev.	Dates	Colour
Quaker Committee on the Aborigines, <i>Protecting the Empire's Humanity</i> and Wellcome Institute	QCA, PEH, WEL	1832/37–1846	Orange
Aborigines Protection Society	APS	1837–1919	Purple
Ethnological Society of London	ESL	1843–1871	Blue
Anthropological Society of London	ASL	1863–1871	Green
Anthropological Institute	AI	1843–1871	Grey

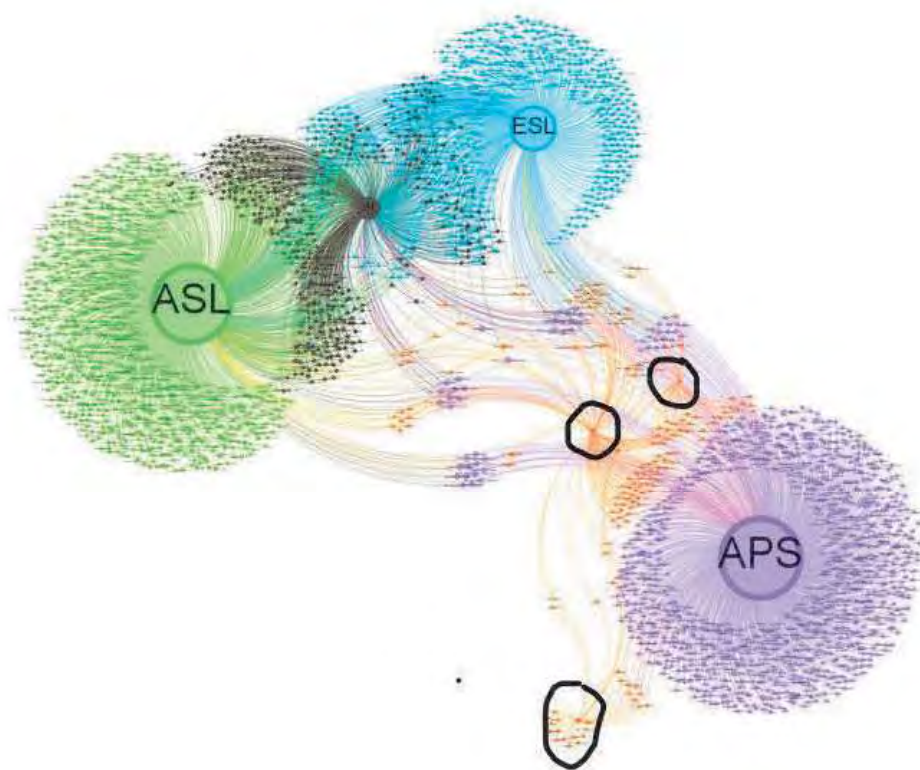


Figure 6.50 The CEDA with 'ZOË' and 'WEL'

6.22.6 The Quaker members of the HOD

Figure 6.52 shows the Quaker families in the HOD grouping. Those Quakers with no or few family members appear in grey (bottom right). The graph illustrates the significant presence of Quaker families in this Case Study.

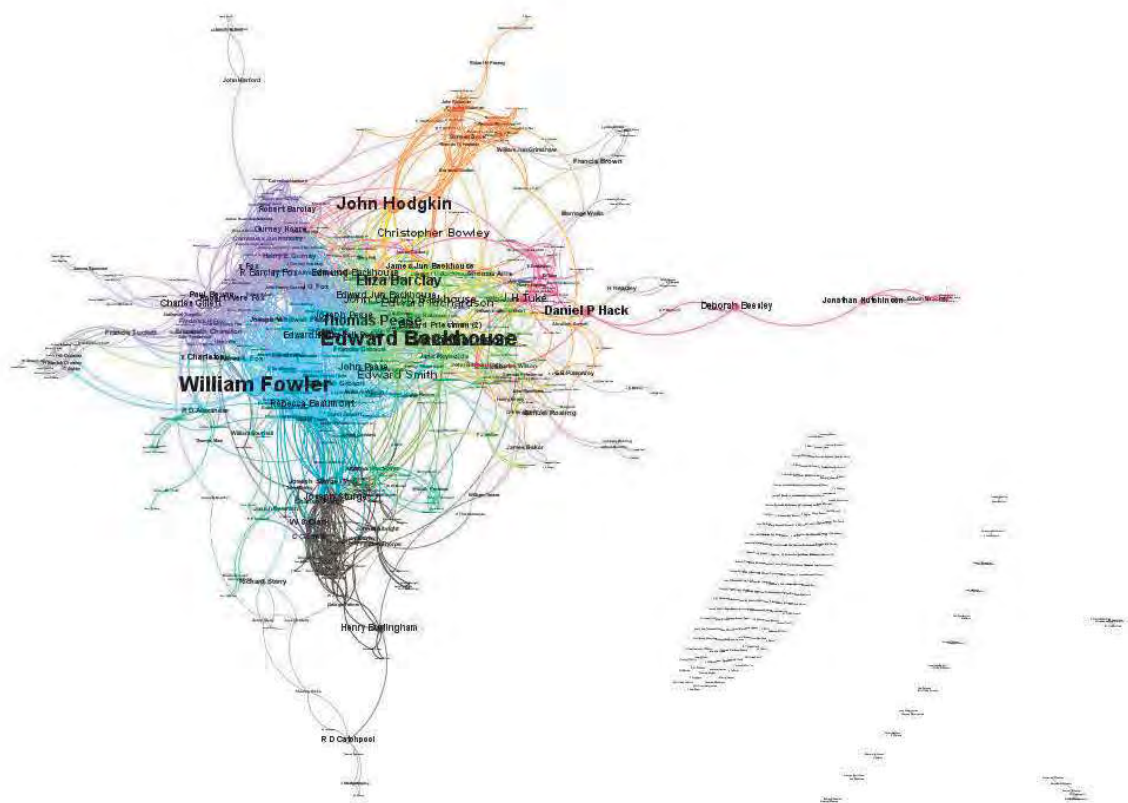


Figure 6.51 Quaker families in the HOD network

6.22.7 The emergence of smaller groups

We can see that both the ZOË and WEL dummy groups are centred by the Force Atlas graph and each connects up all of the main CEDA groups. This is visible confirmation that the Thomas Hodgkin MD network referenced in PEH is well placed and well connected. Other smaller groups that liaise between Hodgkin and the CEDA also become visible in the graph, and these are worthy of further study (see Figure 6.53).

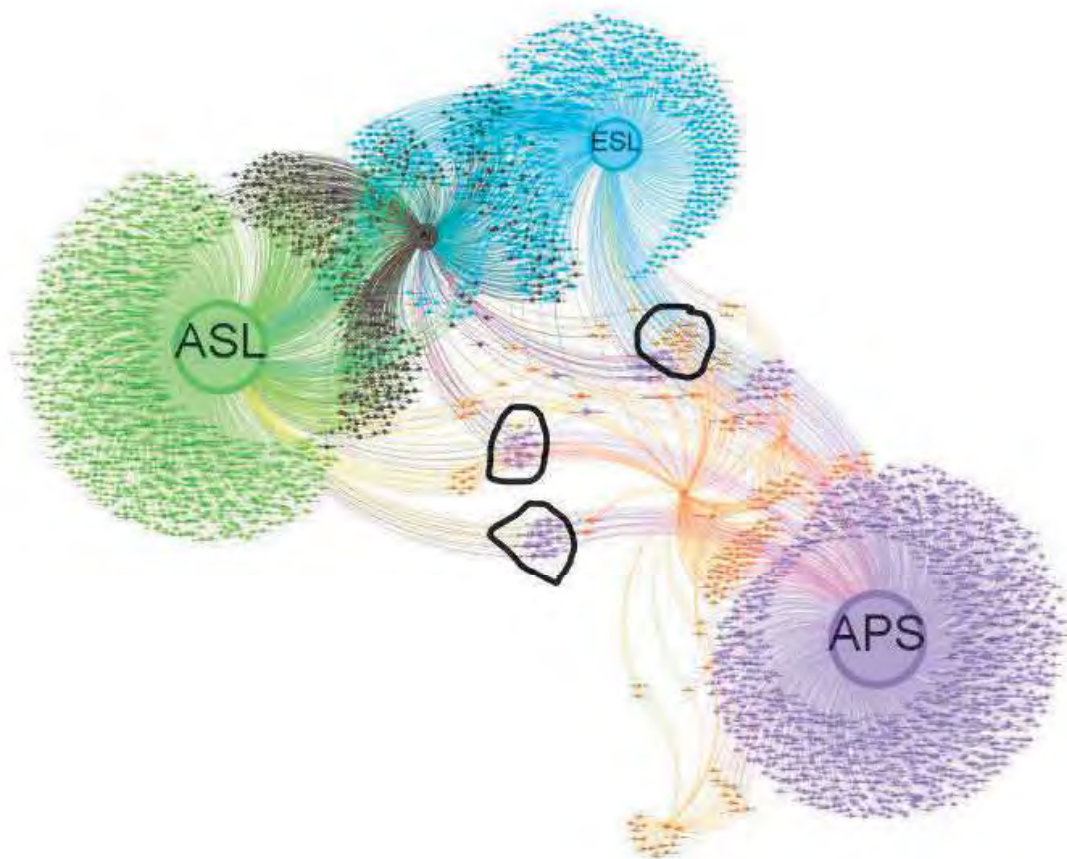


Figure 6.52 The emergence of smaller groups

The presence of the relatively small dummy groups brings Thomas Hodgkin MD to the very centre of the graph; this is impressive given the size of the larger groupings. But it is important to note that Hodgkin does not 'sit' within the community referenced by Laidlaw in PEH or within that suggested by the WEL analysis. Instead he sits to one side, because as much as he is attracted to the PEH and WEL groups, he is 'pulled' away from them through his other connections to the QCA and ESL (see Figure 6.54).

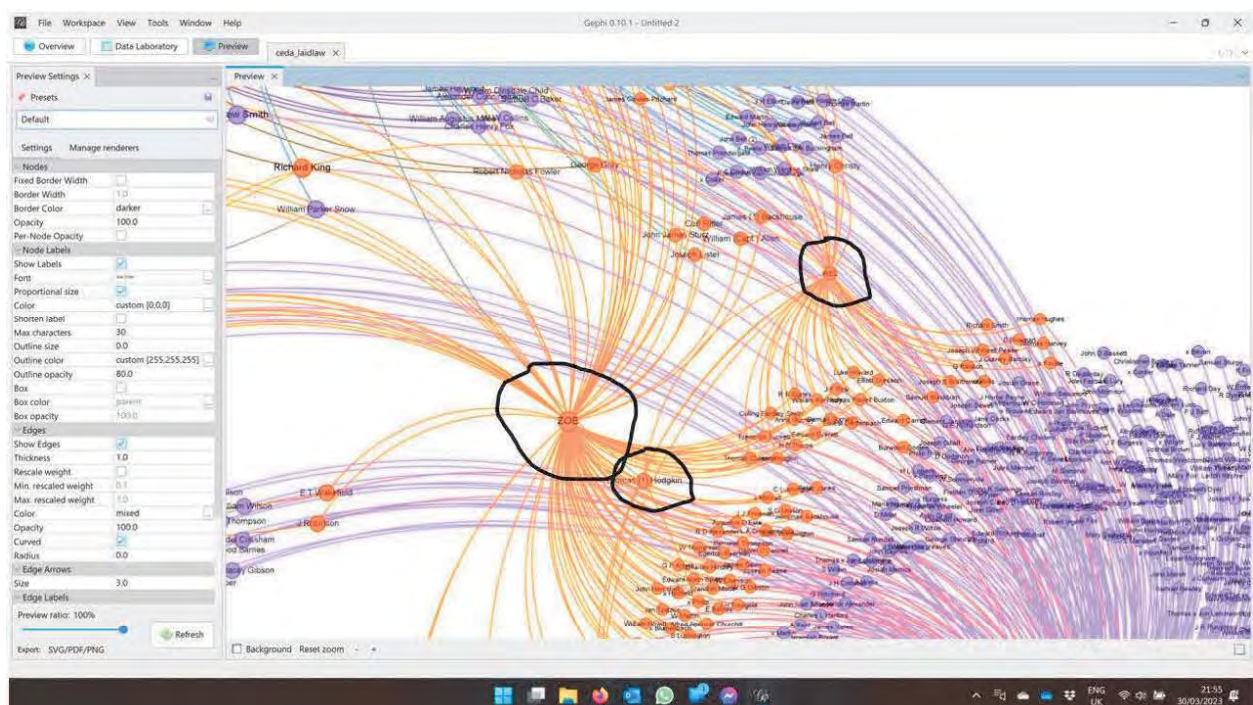


Figure 6.53 Hodgkin apart from 'ZOË' and 'WEL')

6.22.8 Key individuals emerge

PEH includes six (of the fifteen) Quakers who were members of the QCA (see Figure 6.55):

Josiah and William Forster, Robert Howard, Peter Bedford, Joseph Sturge and Robert Alsop (Jnr), and they can be seen in the graph showing networking between the QCA, APS, ZOË

and Thomas Hodgkin MD. This indicates that they have significant roles in the CEDA and are suitable for further study.

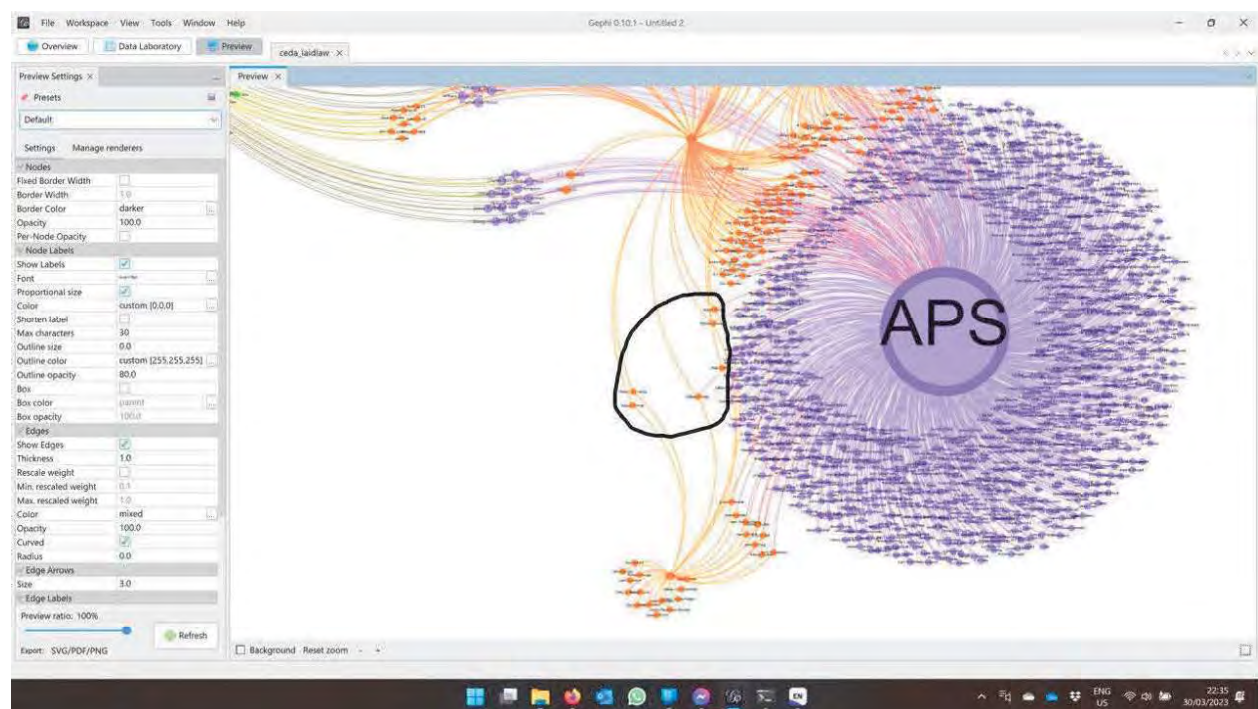


Figure 6.54 The emergence of key individuals

6.22.9 Case Study 3 Conclusions

Question 3 revisited

Can the model reveal the key networking role played by the Quaker Thomas Hodgkin MD (1798–1866) from the beginnings of the CEDA in 1830 up to his death in 1866?

This exercise has shown that although Thomas Hodgkin's networks appear in alphabetical list form in PEH and similarly in WEL, in neither of these sources can Hodgkin's political activist relationships be easily deduced. But by using data analysis and data visualisation technology, the networks argued for in PEH and extracted from WEL can be shown and analysed very clearly.

The HDDT CEDA database comprises the 3000 memberships of CEDA societies concerned firstly with the plight of aborigines, but then quickly afterwards moving towards institution building in the science of ethnology and anthropology, and within all of these groups Thomas Hodgkin has been shown to be a key influencer. Using Gephi topology a much fuller view of the CEDA's political activist relationships and Hodgkin's centrality within them is obtained. Other smaller embedded networking groups that liaise between Hodgkin and the CEDA have been revealed, and these are also worthy of further detailed study.

Although *Protecting the Empire's Humanity* offers throughout the publication a literary description of Hodgkin's networks, only through Gephi topology do we see that Hodgkin himself does not get subsumed into the ZOË or WEL networks; instead he stands aside. This is because, as we can see, his other relationships with both other Quakers and other members of the APS are strong enough to be shown separately from the ZOË or WEL networks. This is because there are others in the CEDA database connected to Hodgkin who do not appear in either PEH or WEL. This finding is important, because it gives good reason to examine Hodgkin's relationships further within both the QCA and the APS (and perhaps also the ESL). Six Quakers who are key networkers who also bridge Hodgkin's concern for the plight of aborigines and wider institution building in the science of anthropology have also been identified in the Gephi graphs.

This Case Study confirms and supports the centrality and importance of Hodgkin's networks as set out in the index to *Protecting the Empire's Humanity* and the Hodgkin Family Papers archive, and invites further scrutiny of those networks because key individuals and small clusters of persons emerge from the visual analysis as key networkers alongside Hodgkin.

6.23 Project Seven review

Project Seven has been added to the comparison table used to assess recent research projects in Chapter 5 (see Table 6.2).

Project Title	Social Networks and Archival Context	Cambridge Group Integrated Census Micro-data	Traces through Time	ResearchSpace	Golden Agents	Project Seven
Project infrastructure regime (Section 5.7.1)	US	UK	UK	UK	EU	N/A
Archivist/ researcher balance (Section 5.7.2)	Both	Research	Archive	Archive	Both	Both
Evidence Based Prosopography (Section 5.7.3)	Mixed	Yes	Mixed	Mixed	Yes	Yes
Data management (Section 5.7.4)	Open	Closed	Closed	Closed	Closed	Open
Genealogical data component (Section 5.7.5)	No	Yes	No	No	Yes	Yes

Provenance (Section 5.7.6)	Database	Primary source	Database	Database	Primary source	Mixed
Technologies and integrated tools (Section 5.7.7)	Yes	Yes	No	Yes	Yes	Yes
Usability and user interventions (Section 5.7.8)	No	No	No	Yes	Yes	Yes
Visualisation affordance (Section 5.7.9)	yes	No	No	Yes	Yes	Yes
On-going support (Section 5.7.10)	No	No	No	No	No	No

Table 6.5 Project comparison table

6.23.1 Project infrastructure

P7 infrastructural support came from the University of Birmingham, which provided dedicated RSE support in two phases: (1) to clean data and build and populate the database and (2) to integrate JNB and Gephi into the HDDT. Project steering was provided by the IHR. Other training needs were met by LinkedIn, StackOverflow and YouTube by self-searching for help when it was needed. This low level of support did not adversely impact the project, and perhaps it gave the P7 author a level of freedom greater than there would have been had more expertise been assigned to the project. On the other hand, the project could have

used more EBPD, which is generally available, for instance on Quaker occupations. Elizabeth Careless had compiled a database of EBP on all members of the ESL in 1974.⁵⁶⁷

6.23.2 Archivist/researcher balance

The RAI provided core data for the P7 project in the form of an extract from an old non-standardised metadata finding aid. The Genealogist identified Quakers in the data and then provided family relationships by extracting them from a Gedcom file which he owns. Other data was extracted manually from microfilm (which was a representation) and manuscripts (which were primary sources). The database records could have been much richer in content, as there is an abundance of data digitally available. The object of building and using the HDDT was only to demonstrate the utility of DH modelling and data handling – data collection stopped once sufficient data had been collected. The database could be easily expanded but to do so would require (a) substantially more resources and (2) a new rationale for setting the boundary to the data scope. Therefore, the data, while being appropriate to this project, should not be regarded as a complete dataset.

6.23.3 Evidence Based Prosopography (metadata versus primary source)

The P7 project used a mixture of metadata: primary and secondary sources. The blended dataset worked well as an integrated package, but it was difficult to separate the two data

⁵⁶⁷ There are many suitable data sources that might provide further EBPD on the members of the CEDA. For example, (Careless 1974): the database no longer exists but outputs from it are appended to the research paper; (E. Milligan 2009); The papers of the British Association for the Advancement of Science at the Bodleian archives; and the databases of the Royal College of Surgeons and the Royal Geological Society.

qualities at the output end of the project. Because the input data from all sources arrived as CSV sheets, there was a uniformity which disguised the underlying variability in the quality of the sources. This uniformity, and the data conformity which arose when the combined data was reproduced in the database, lost sight of input data variability. This is a concern for all DH projects that start with data of variable quality and end up with a blended, clean and tidy database.

The EBPB was collected from lists of names in APS publications (see Project Seven Report Figure 7.4.10). These names can be identified with reference to the date of the meeting at which the membership secretary recorded the names, repetition of the name in other APS meetings and the genealogical data provided by the Genealogist on the Quakers present in the CEDA database. Had the NAI-UID system been in place it would have been possible to append the NAI-UID UID and for the NAI-UID system to record that this name had been sighted in this source and in this digital record (the Project Seven Report). This would, in turn, have enabled a later researcher to challenge the allocation of the name and suggest alternatives. Without an NAI-UID system it is not possible to accumulate and organise EBPB and to be able to (1) track EBPB contained in the digital references back to the source of the EBPB in the physical source and (2) enable researchers to browse and consider both the location of all sources containing the EBPB and the possibly many digital references made to it.

6.23.4 Data management

The metadata was old (pre-digitisation) and needed considerable cleaning. Because the input data was variously structured and relatively low volume, and it arrived as CSV sheets, Excel was chosen to clean the data. It would have been possible to use algorithms, but doing so would still have required data manipulation to take place before cleaning with algorithms could have been utilised. I, as a 'Independent Researcher', was familiar with Excel but not algorithms, so the decision to use Excel was as much based on my skill set and the burden of learning already committed to in the other technologies that made up the HDDT. Using VSC and GitHub helped greatly in managing and systematising the data management process, and auditing and interventions were built into the project scheme. Both were frequently needed and used.

6.23.5 Genealogy

The desire to incorporate family relationships into the P7 project came late, after the HTTD had been built. It is not certain whether data from Gedcom files⁵⁶⁸ could be integrated seamlessly into an HDDT built around an SQLite relational database, but in any case it was beyond the scope and skill set of the project. Nevertheless, a work-around was devised to use simplified family relationship indicators and to apply them person to person in the database. The P7 project successfully applied the work-around to enable the HDDT to present family relationship between the Quakers in the database. It would be possible to

⁵⁶⁸ 'A GEDCOM file consists of a header section, records, and a trailer section. Within these sections, records represent people (INDI record), families (FAM records), sources of information (SOUR records), and other miscellaneous records, including notes. Every line of a GEDCOM file begins with a level number where all top-level records (HEAD, TRLR, SUBN, and each INDI, FAM, OBJE, NOTE, REPO, SOUR, and SUBM) begin with a line with level 0, while other level numbers are positive integers. Although it is possible to write a GEDCOM file by hand, the format was designed to be used with software and thus is not especially human-friendly.' <https://en.wikipedia.org/wiki/GEDCOM> (Accessed 26 January 2025).

extend the HDDT data to include the family relationships of all CEDA members, but this would extend the timeline of the project considerably. This would be a questionable exercise given that in the HDDT family relationship data is only useful as a support for the HDDT's main objectives.

6.23.6 Integrated tools

The P7 project integrated VSC, GitHub, Sqlite, JNB and Gephi free affordances seamlessly, and the HDDT worked robustly throughout. The HDDT is an example of a research model that can be built by a Independent Researcher (with technical support) and is only one of a multitude of possible technological configurations, with more arriving every day. This itself is a concern because each affordance requires training and skills acquisition, which is a considerable investment for a Independent Researcher in affordances which have perhaps a short lifespan in terms of popularity and online support.

6.23.7 User interventions

Each of the components of the HDDT allowed user interventions, and these took place frequently to locate problem data and to remedy poor data inputs. As data travelled through the HDDT pipeline it was rendered differently and so errors resurfaced in data cleaned in Excel when they were rendered in SQL, JNB or Gephi. Data cleaning was necessary throughout the data pipeline and this would have been a prohibitive problem if user interventions had been absent or inadequate.

6.23.8 Visualisations

The HDDT used only Gephi for network visualisations. Gephi requires considerable training and experience to use. The platform is used in different ways by different researcher communities, and each has their own preferred methods and practices. User training in humanities data is poor. It was very easy to spoil a visualisation and impossible to walk back to an earlier step – instead, one had simply to start again. The number of options at each stage is large and there are several stages. It was easy to lose track of all of the choices made to create a satisfactory network visualisation and hard to remember how to repeat the same selections in the same order for the next visualisation. A standard presentation of network graphs was desirable and so Gephi worked best by rigidly following a strict script at all times. This discourages the use of Gephi to explore alternative options. Gephi calls for experienced operators and it takes time for a Independent Researcher to acquire this experience.

6.23.9 Ongoing support

P7 is a Independent Researcher project with no data to offer to a data archive. It will be abandoned at project completion. It is the account of working as a Independent Researcher and the experience in developing, building, populating and using an HTTD (which is easily reproduceable using a variety of technologies) which will be made available for the benefit of future researchers.

6.24 Project Seven conclusion

Project Seven was a success as a stand-alone short-life project and as an example of how Independent Researchers can make good use of technologies that perform at least as well as those used in more expensive DH projects carried out either in stand-alone research hubs (US-style infrastructure) or centrally supported research projects (EU-style infrastructure). The HDDT works well as a research tool and the P7 Report is a good example of the quality of research practice that is available to the Independent Researcher. Project Seven brought to light many of the concerns a researcher would have about using digital affordances – digital records where the source is lost, the use of metadata (which is a secondary source) as a surrogate for sources, mixing metadata with representative digital records and poor provenance tracking. For the P7 author the most important benefit in building the HDDT and using it to answer research questions has been the opportunity, through practice, to question every aspect of DH research and to go on to expose weaknesses in the entire DH process, and then find enduring improvements to DH practice.

Chapter 7 Conclusion

7.1 Reflection

7.1.1 From information in sources to EBPI and EBPD – the challenge

Citizens of today's world have many unique identifiers (UIDs) attached to the information about their lives – tax numbers, National Insurance numbers, National Health numbers, driving licence numbers, bank account numbers, etc. Using these numbers it is possible to locate, verify and organise information about current human lives. Of course, we do not do this, because to do so would breach data protection laws and privacy conventions. But what of the dead? They are not bound by data protection laws, but then again, neither do they yet have nationally structured arrays of UIDs that can be used to systematise and organise information about their lives. This thesis shows that we can retro-fit UIDs and then use these to help us better understand individual and collective PHL. This is the Evidence Based Prosopography and NAI-UID system. This thesis has shown how EBP can become the systematic organising concept of information on PHL (EBPI) by representing it as data (EBPD). Adoption of the EBP system will not impose an unnatural order on important information on PHL, but instead EBP exploits the natural order already present in information on PHL in order to direct the structure of the EBP system.

The critical defining feature of Records of information is their uniqueness, fragility and uncertain durability. The first organisational dilemma of Records is the impossibility of classifying them, because they arise from the past in a vast array of widely varied forms,

where often the form itself is intrinsic to the cultural meaning and value of the Record. The second organisational dilemma is that the remains of the physical world of the past are chaotic, messy and beyond universal organisation. The digital world of data is not the same as the past world of Records – it is (arguably) less fragile and more durable, it is naturally organised by classes of things, in rows and columns, and the data is therefore necessarily conformative and structured. In this sense the digital world offers structure and organisation to the world of information in Records. The difficulty is how to strongly connect information to data, and how to make those connections durable.

The critical defining feature of the digital representations of information in Records is that EBPD must never become knowingly unattached from the EBPI it represents. Data can represent or reproduce information in Records to varying extents and in various ways, but data cannot not replace information in Records. In the digital world, the evidences of PHL must consist of faithful representations of Records and of the information contained in those Records (that still exist). Over time, the digital representations of many Records may be all we have left of the evidences of PHL as Records fade away. Each Record is unique (although there can be many instances of a given type of Record) and because Records are physical things, they constantly require care and attention. They are also widely distributed around the globe and are hard to find, even with the help of digitised finding aids. Digital representations of Records and the information they contain have the ability to emulate by representative verisimilitude the sources and information themselves, in a durable, universally accessible, virtual environment – and it is in this aspect of digital manifestation that the importance of fixity arises.

Digital representations can overcome the problems that Records cannot shake off: fragility, cost (of housing) and difficulty of access. But digital representations bring with them their own problems, particularly accreditation, fixity, affixedness and authority.⁵⁶⁹ These problems arise because Records can be moved, merged with other Records or lost, and digital representations can be copied (sometimes extensively), altered, edited or summarised as metadata. All of this makes it very likely, especially over time, that the connection between EBPI in Records and their EBP digital representations will be broken or lost. This special relationship, between Records and their digital representations, and the need to systematise and order those relationships, is the central concern of this thesis.

All of this makes it likely, especially over time, that the connection between EBPI in Records and their EBP digital representations will be frequently broken or lost.

This information to data concern is a problem that data science can address, because data science uses established information science methodologies and demonstrable practices appropriate to the husbanding of relationships as information becomes data.

The current state of DH, with its inheritance of a tradition of research interest in a select few human lives and a pull to preserve and reproduce pre-digital structures in the academy, does not yet fully recognise the importance of data science in the study of PHL, and the requirement for all researchers, those in academies and those outside academies, to be equally supported and their need for all EBPI to be equally accessible.

⁵⁶⁹ Accreditation: is this a certified digital representation? Fixity: has this been edited, altered, 'improved' in any way? Affixedness: can this digital representation be traced back to the source it purports to represent? Authority: this representation may have been reproduced many times, so by whose authority is it offered for research (the archive, the researcher, the person who obtained it)?

7.1.2 Digitising the study of PHL – current practice

The digital world is very young and so far a large part of the digitising effort in the humanities that is concerned with information on PHL has been focused on making and sharing discrete digitisation programmes of special collections⁵⁷⁰ and the digitisation of nearly all archival Records as metadata – a global effort by many now unknown archivists, who often painstakingly created the sources of their metadata many years ago, by examining and studying the Records held in their academic collections. And to do both of those digitising activities, they relied on the work of huge numbers of GLAMS specialists. Working over several generations, GLAMS specialists have explained, interpreted, collected and connected up as metadata digital references to both Records and in some cases the information found in those Records – and the digital affordances they have created continue to be digitised apace.

The original finding aids that underpinned these digitising programmes, paradoxically, are themselves physical things, made before digitisation, and so as well as being metadata they are themselves also Records and subject to the same degenerative concerns that all Records face. In the past, archival Record makers used their specialist knowledge and their collector enthusiasm to create physical finding aids, unconcerned that the physical Records they were making would one day be used as data inputs in a global enterprise to digitise and, through metadata, link up dispersed collections and archives into new digital archives. Paper and

⁵⁷⁰ For example TNA's First World War 100 digital collection, <https://www.nationalarchives.gov.uk/first-world-war> (Accessed 2 February 2025) and The Proceedings of the Old Bailey Online, <https://www.oldbaileyonline.org> (Accessed 2 February 2025).

early stand-alone database finding aids are now rapidly becoming large-scale and sometimes interoperable, open digital databases.

This thesis has shown that recently some archives have set about also using these databases to construct research-assistive technological models, as if to point the way for researchers in general to work with and within the new digital archive.⁵⁷¹ In these digitising efforts, the archival community has embraced the designing of digital standards (such as CIDOC-CRM), developed complex methodologies (such as the Golden Agents project) and utilised (often borrowed) restrictive technologies as tools (the SNAC project), in order to find, clean, assimilate and create metadata affordances as assistive technologies. As a consequence, we have made efforts to clean up, systematise and improve past archival Record keeping mores, and we have learned how to innovate in sharing and interconnecting the resulting metadata. However, this has sometimes resulted in the digitising of unstructured or messy data, by forcing relatively incompatible physical information into structured digital approximations. The immense effort of digitisation in the humanities is to be rightly applauded, because while the digital archival affordances do not intentionally represent Records, they more effectively and more comprehensively point the way to Records. But digital archival records have their limitations, especially for researchers keen to work with information as EBPI.

The digitisation and linking up of old GLAMS records form only the first wave of the digitisation of the evidences of PHL and like all first efforts they are a good and important stepping stone to the next stage. However, to move to the next stage we first must

⁵⁷¹ Archives have responded to the call from researchers for access to sources and have reasonably turned to their new digital affordances to create assistive offerings to researchers that go some way towards meeting researchers' needs.

recognise and accept that the digital representations made in the first wave are, of course, all secondary sources. They are not direct representations of the Records they refer to or of the informations contained in those Records.⁵⁷² They are instead subjective and limited descriptions of them.⁵⁷³ While they are of great help to the archivist, they are of arguably less use to researchers whose primary interest is in the Records themselves (unless the researchers' interest is in the digitisation process itself).⁵⁷⁴

7.1.3 Digitising the study of PHL – new practice

The next phase of the DH enterprise must be the comprehensive digitisation of prosopographical information contained in Records. Prosopographical information satisfies the need to respect the original information by only digitally extracting from the Record the information that can be copied to data unmediated, as it is, EBPI to EBPD. Prosopographical data fits the structural needs of databases because it consists of person names and

⁵⁷² They are unattributed, unstructured accounts forced into a digital conformance. Furthermore, these records are often limited to the lives and deeds of significant persons or more ordinary persons caught up in events of national or cultural prestige.

⁵⁷³ A study of users of finding aids (FA) revealed that users were preoccupied with getting access to sources. See (Freund and Toms 2016, 1005): 'Scope and Content: This clearly emerged as the most heavily used and important element for understanding what is contained in a collection: "the meat and potatoes" (H02) of the FA, and the "most useful section, because it tells you if what you're looking for is there" (01 1). It provides an overview and helps filter materials: "the scope and content would give you a good overview of what sort of things you should expect. I think that would be the premier thing, the first thing. And sometimes even the only thing, because ... you could often rule out a whole set of files" (H04). However, participants were aware that it was not useful as a source of specific information: "scope and content isn't particularly valuable because I might be interested in a specific church or a specific building and it doesn't really tell you that" (H1 5).'

⁵⁷⁴ 'As we look through archival history, we will see how the characteristics of finding aids as tools consistently colored how archivists understood access and defined the emergence of descriptive standards, often in limiting or negative ways. By the 1990s, finding aids became actively detrimental as they framed the creation of the Encoded Archival Description (EAD) standard. It is more appropriate to describe EAD as encoding finding aids rather than encoding archival description, as the standard preserved the remains of the entire messy history of finding aids within its tags. While online finding aids greatly improve access to archival materials overall, the remnants of finding aids past undermine their effectiveness in multiple ways' (Wiedeman 2019, 382).

attributes which can be ordered into rows and columns; it does not require transformation when it moves from EBPI to EBP. Therefore, prosopographical information forms the best bridge between the world of Records and information, and the digital world of the database and its data. Furthermore, prosopographical information can be developed extensively to build family and other relationships, joining up PHL data by a natural relational characteristic embedded in the information itself. This enables information and Records of all PHL to be interconnected to take full advantage of the features of the Semantic Web.

In using genealogy to structure, order and interconnect the evidences of PHL, we simply follow the natural and universal structure of the information itself – the family relationships that can be attributed to named persons. It is this natural structure on which a national system of authority for PHL can be built: the NAI-UID system. If EBP and the NAI-UID system are implemented, the next phase of digitisation in the humanities will then go a long way towards fulfilling Salvatore Spina's vision:

Our future lies in the archives and their heritage, through which we write history. But, if the historical archives represent the 'databases' of the past, the 'future of the past' lies in the greatest database that history has ever created: The genetic heritage, which is, on the one hand, the 'a priori' structure of the performance of man's action, and, on the other hand, a natural archive that lies inside every one of us. (<https://www.timemachine.eu/ambassadors/salvatore-spina>, accessed 01 February 2025)

Structuring, ordering and digitising the evidences of PHL in the form of EBP is not only possible but also desirable.⁵⁷⁵ To do this, DH must begin the task of linking prosopographical

⁵⁷⁵ 'Today's student of literature must be adept at reading and gathering evidence from individual texts and equally adept at accessing and mining digital-text repositories. And mining here really is the key word in context. Literary scholars must learn to go beyond search. In search we go after a single nugget, carefully panning in the river of prose. At the risk of giving offense to the environmentalists, what is needed now is the literary equivalent of open-pit mining or hydraulicking. We are proficient at electronic search and comfortable searching digital collections for some piece of evidence to support an argument, but the sheer amount of data now available makes search ineffectual as a means of evidence gathering. Close reading, digital searching, will continue to reveal nuggets, while the deeper veins lie buried beneath the mass of gravel layered above. What are required are methods for aggregating and making sense out of both the nuggets and the tailings' (Jockers 2013, 9).

information in Records to its Representative Data. Doing so will require the skills of historians, genealogists, technologists and data scientists to be brought together, and will involve reaching beyond academia to work with crowdsourcing endeavours.⁵⁷⁶ Historians bring with them the skills to judge what is, and what is not, prosopographical information. Genealogists bring with them the skills of using judgement in the attachment (matching) of digital representations to Records. This is a complex process because information in Records is often messy and sometimes insufficient to make a straightforward allocation. It can also sometimes be capable of multiple attachments. For this reason the NAI-UID system is recommended because it records and makes available to scrutiny the matching judgements made.

EBPD can currently be extracted from EBPI in Records by researchers piecemeal (see Section 5.5), but without the integrated overlay of the EBP-NAI system, localised research practices will risk being unintegrated and uninteroperable practices.

The commitment to move to a more shared and more integrated practice requires a comprehensive data service infrastructure such as the EBP NAI-UID system. A system which

⁵⁷⁶ 'In the interdisciplinary field as well, this evolution has more than ever reinforced what we could classify as silent change in history and in the humanities. In the case of GIS, for example, various approaches to linguistics and literature have been explored. This can be seen, e.g., in the recent works of Ian Gregory, a specialist in historical GIS, in collaboration with colleagues in those areas. Archaeology, classical studies, and computational science have equally worked together as in the case of Google Ancient Places project, for instance. Urban history, literature and methods from ecology has come together through the use of digital texts and digital tools. Other projects gather specialists from areas as diverse as geography, history, economics, computational science, and physics as with the project Water, Road & Rail. This crescendo of interdisciplinary work – whether due to research projects crossing knowledge from various disciplines, whether due to the use of certain methodologies or in order to reach certain results, the historian is now more aware that it is necessary to resort to specific collaborations with other specialists – is today an undeniable reality. Perhaps attentive observation to the indexes of scientific magazines, an accounting of the number of authors per article, and the identification of their particular curriculums would allow us to confirm that the work of the historian is no longer that of a tireless archive researcher who continues, in an insular way and over many years, on a single-themed investigation, or the work of someone who will singlehandedly take credit for the results. Whether we like it or not, this change, silently, has established itself as a revolutionary (it is difficult to escape the term!) element of the historian's work and, in my opinion, it represents added value to the diversification, democratization, and deepening of historical knowledge' (Rocha 2009, 6-7).

structures how information on PHL (EBPI) becomes EBP and enables EBP to be researched locally and also interconnected at a national level. Importantly, and to recognise the universal characteristic of messy data, the NAI-UID is a permissive system, allowing information contained in an instance in a Record to be allocated by different researchers to different NAI-UID index records, based on individual judgement. Technologists bring with them skills in making choices of appropriate research technologies for each stage of the research process, integrating technologies into seamless research affordances (such as the HDDT) and scripting code efficiently and interoperability. Data scientists bring skills such as design and an overall understanding of the entire process from information on PHL to research outputs, including the implementation of standards and the critical assessment of research processes and practices (see the seven stages of the P7 project, Section 6.1).

7.2 Returning to the research questions

1 What is Evidence Based Prosopography and the National Authority Index system?

Are they the way forward for digitisation in the humanities?

This thesis fully described Evidence Based Prosopography and the National Archival Index system in Chapter 2. A balance has been struck between describing a fully developed system and presenting an outline, because EBP and the NAI-UID system can only be established by the consideration of many people, including a wide range of expertise, and this will take time. Instead, what has been presented here is an indicative outline of the EBP and NAI-UID system sufficient to attract the attention of others. Project Seven (Chapter 6) has shown how EBP can be utilised in a practical exercise demonstrated through three Case Studies.

Section 5.5 has shown a similar contemporaneous research project using EBP in the Huygens Institute and, noting that it is fully compliant with the NAI-UID system, recommends the Golden Agents project as an exemplar of research practice compliant with the EBP NAI-UID system.

2 Is infrastructure provision in the Digital Humanities sufficient to take up the national enterprise of the digitisation of Evidence Based Prosopography?

The thesis has reviewed infrastructure provision on a regional basis in the US, EU and UK, and found that although there are regional differences in approach (top down in the EU, bottom up in the US and mixed in the UK), the levels of provision are nonetheless broadly similar at archival and researcher levels. This is evidenced in international standards integration across multiple archives and across the US, EU and UK. There is clear evidence of cross-continental working at the researcher level (SNAC and ResearchSpace). This indicates that EBP could be universally adopted, embraced and nurtured by all three infrastructural support systems. Because EBP utilises and makes accessible information structures naturally found in Records, it is unlikely that adopting EBP would adversely impact the further development of infrastructure provision in DH. EBP utilises the structure of familial relationships present in the Records of PHL in the archive, and universally present in genealogical affordances, to build the EBP system. This indicates that genealogical infrastructures and practices will support the development of the EBP system.

3 Are digitised finding aids a good bridge to the Records and the informations contained in them that researchers are interested in?

Considerable efforts have been made in the archives to digitise archival finding aids (Chapter 4) and to use the resulting digital metadata records not only as improved finding

aids, but also occasionally to undertake institutional research into PHL (Chapter 5). Standards harmonisation, technological developments (especially in the areas of interoperability and data management) and infrastructure developments at international, national and institutional levels in the archive (Chapter 3) have all resulted in digitisation becoming deeply embedded in archives and institutions. Archivists in the US, EU and UK have digital skills and are practised in working with digitally with Records; they are ready for a new phase of digitisation. Digital archival records are comprehensive and while they are in themselves secondary sources, they are the best affordance in the archive to point the way to Records and they are a central part of building up the EBP system. Similarly, the global virtual archive now being built is an important part of the EBP system because it provides an umbrella infrastructure that will enable Records and the informations contained in them to be matched to the EBP data system when it is adopted in the US and the EU. Digital archival practice in the management of archival records will also provide expertise in the development of the EBP system.

4. How successfully have recent large-scale research projects into Past Human Lives used Evidence Based Prosopography?
5. This thesis has shown that recent projects that rely exclusively on digital archival records can struggle to deliver a satisfactory offering to researchers (Chapter 5), although they do more often provide a satisfactory offering to Records management and organisation activity in the archive (Chapter 4).
6. Projects that rely on EBP often either fail to adequately show the relationships between the records themselves⁵⁷⁷ or to make publicly available the precise

⁵⁷⁷ In the case of The Cambridge Group this failing is contractual – the commercial provider prohibits free distribution of the linkages.

location of referenced instances of EBP in the related Records (Section 5.2.1). The Golden Agents project makes full use of the EBP system at the project level, but, because there is as yet no NAI in the Netherlands, it cannot do so at the PHL person name level. In all of the cases considered here, the EBP and NAI-UID system could have been incorporated into the project, and it could have significantly improved the future findability and usability of each project's data. This thesis author is concerned that without the EBP system, and with the natural retirement of project teams, the future usability of data is in doubt.

5 How successfully has the small-scale Project Seven research project into Past Human Lives used Evidence Based Prosopography?

The Project Seven development undertaken by me has shown that Independent Researchers can today acquire the digital skills and technologies (at low or zero cost) to perform detailed and complex analyses into PHL at levels of project performance that closely match the performance of digital research projects done by larger institutions, and at much greater cost. P7 has also made clear the limitations of digitally researching PHL, and that the most important limitations remains as they were before digitisation:

- poor access to Records
- messy data
- incomplete data
- unstructured and disordered information
- lack of digital access to basic prosopographical information contained in Records

P7 would have been more successful if the EBP system had been adopted at the archives where information was collected and if these had been linked to an NAI. The P7 project would also have been readily scalable and made more interoperable. Had the EBP NAI-UID system been fully operational, then P7 would have benefited in the following ways:

- P7's RAI-UID would have been matched to the national NAI-UID, allowing all of the person records in the project to be discoverable in a single national-level enquiry.
- Information extracted from manuscript Records would have the location of the Records identified through the AAI-UID Index.⁵⁷⁸
- Information extracted from genealogical Records would have the location of the sources identified through a GAI-UID Index.
- The Lone Researcher's credentials and accreditation would have been discoverable.
- P7 would have been registered as a research project through the researcher's sponsoring institution.

7.3 A data science perspective on the study of Past Human Lives

Chapter 1 finished with the observation that, after completion of the work of this thesis, a view from a data science perspective should be offered concerning the relatively slow take-up of the digital in the humanities. This thesis has shown that in large part this is because many of the digital affordances created so far, whether based in the archive or genealogy,

⁵⁷⁸ Microfilm, from which data was extracted and which then became the backbone of the combined dataset, would have been identified as not attached to its source. In P7's case it would not have been possible to also identify the location of the sources, because they are lost. P7 took the view that in this project the microfilm is the source.

frequently do not meet the basic need of researchers, which is to access and work digitally with data that references and represents physical information on Past Human Lives, wherever that is found(Section 2.10). It can be argued that take-up of digital methods and practices should increase significantly if the EBP system were in place, because it would provide the necessary infrastructure support for research. Researchers would be able to trust and use digital representations because (1) they would be digitally findable, (2) their provenance could be checked by reference to the location (institution and collection) of the respective Records, (3) the credentials of the EBP provider could be checked (whether the provider is an archive or another researcher), (4) their own research could be linked to both other research projects and to the individual Records researched, and (5) past users of the same data would be findable. In respect of technical competence, this could be provided by suitable institutions. This thesis has championed the role of the Independent Researcher who wishes to study PHL (Chapter 6). There is a mismatch between the large number of members of the general public who routinely work online as amateur family historians on genealogical platforms, working in a disciplined manner with Records and Representative Data, and the relatively few academic researchers working with archival Records. If lone academic researchers had at their disposal well-supported infrastructures that spanned both the archive and genealogy, then the barriers between the academy and genealogy would fall away.

7.3.1 Generative Artificial Intelligence and Past Human Lives

At the time of writing generative artificial intelligence (AI) has emerged and is rapidly being deployed across the internet – swiftly in the commercial sector and more cautiously in the

academic sector.⁵⁷⁹ The current novelty of AI means that it is today unclear to what extent AI will become a part of the everyday and academic world.⁵⁸⁰ Therefore, it is too early to make here anything but an intuitive guess how AI might impact the methodology and practices for study of past human lives set out here. However some early indicators (widening the appeal of research to include non-academics and improved name identification through large language model learning in its development) suggest that AI might become a major accelerator to the study of past human lives once the Evidence Based Prosopographical system has been developed. AI (as it is currently structured in the generative data field and as it applies in this thesis) comprises in digitally finding, prioritising, organising and analysing data, in order to produce a written summary of the selected data in patterns of common speech. It is anticipated that AI will also be able to make both predictions and perform routine actions based on data trawling of the internet and user interactions. To do both of these AI needs both data and user instructions, and both of these must be capable of rational and meaningful organisation. But, if the object is the study of information embedded in physical sources then that information which is not digitised will remain invisible to digital enquiry and thus AI. The EBP system finds and reveals structured and organisable data (on past human lives) and makes it digitally accessible. By

⁵⁷⁹ 'Generative AI, short for Generative Artificial Intelligence, is an exciting subfield of artificial intelligence that focuses on developing systems capable of autonomously generating new and creative content. It enables machines to go beyond traditional tasks like classification and prediction and venture into the realm of imagination and creation. By leveraging deep learning techniques and generative models, these systems can produce novel outputs, such as images, music, text, and more, that closely resemble human-generated content. (Ramdurai and Adhithya 2023, 1)

⁵⁸⁰ For a cautious initial assessment embracing social and political concerns alongside technological advances, see Henrik Skaug Sætra - 'Generative AI has taken the world by storm, kicked off for real by ChatGPT and quickly followed by further development and the release of GPT-4 and similar models from OpenAI's competitors. The street has most certainly found its use for generative artificial intelligence (AI), and there is no longer much point in discussing whether generative AI will be influential. It will, and what remains to be discussed is how influential it will be, and what potential harms arise when we use AI to generate text and other forms of content. Technological change entails societal change, and we must always endeavor to ask how new technologies shapes, engenders, or potentially erodes the "good society". In this sense, Generative AI is another instance of politically and culturally disruptive autonomous technology' (Sætra 2023, 1)

using the NAI system the distribution of that data amongst the physical sources can be found and uncertainties in its disambiguation when referenced as data noted. Thus evidences of past human lives can be presented to search engines in an efficient and managed way. For this reason it can be expected that once the EBP system is up and running, then AI may be able to take advantage of it and make currently invisible information on past human lives more rapidly findable and thereby offer it to researchers to study. The possible application of AI to the study of past human lives might help - because through the EBP system data on past human lives wherever it can be found will be processed in a logical and universally structured form readily accessible in the Semantic web to emergent AI applications.

7.4 Recommendations

The Evidence Based Prosopographical system described and proposed in this thesis is a national system which must be implemented at national level. The building bricks of the system are already in place in the archive and in genealogical platforms. GLAMS archival records (with person name as the primary key) have UIDs allocated and the archival records already point to the Records researchers are interested in – and in some cases also to the incidences of prosopographical data in Records. CIDOC-CRM and other standards in digital archival records already recognise person names and family relationships (see Section 4.2.2). Genealogy platforms have robust systems such as GEDCOM, systematic and embedded data matching capabilities and also the ability for multiple users to make, record

and share independent matching decisions. The US Name Authority Cooperative Project (NACO) and the work of the General Records Office (GRO) in UK National Authority Indexing offer models as a basis for study. At the archive level, archives with digital catalogues and finding aids already have an archival name index in place, although the UIDs are currently hidden.

A national working group should be established to determine if the Evidence Based Prosopographical and National Authority Index system should be adopted in the UK. This would include:

- How archival and genealogical platform index UIDs can be linked to a National Authority Index.
- How genealogical platforms provide a family relationships building facility for GEDCOM, because this could be a model for the development of the EBP system.
- How genealogical platforms preserve the integrity of digital records – fixity, affixedness and provenance – because this could be a model for the design features of the EBP system.
- How genealogical platforms discipline by design the public use of their platforms to protect the integrity of the service, and how they record and share researcher decisions and choices.
- How archival standards could be modified by extension to include the recording of person names present in Records.

7.4.1 The HDDT recommendations

A reading of the P7 report (Appendix 7) shows that this researcher made a considerable investment of time and effort in the acquisition of digital skills. The HDDT was designed and built in 2022, and it is already technologically out of date (in 2025), but the design process and the build and operation of the HDDT use transferable skills. It is these skills that endure. A new HDDT built by this researcher today would use different technologies, but the time and effort to build a new version, and the confidence to build it, would be greatly improved. The thesis shows that because of a lack of familiarity with building such affordances, the Independent Researcher needs considerable support the first time round. Thereafter, support needs diminish. Support must come from an accredited institution, be focused on the specific needs of the researcher and their project, and be in place for all of the development phase of the project. Other assistance may be useful, such as short online learning modules, but they can cause a great deal of confusion because they are generic in nature and often poorly presented, so relying on short online courses can result in a reluctance to persevere.

Institutions should consider offering on-going researcher support to Independent Researchers to develop and establish digital research projects into PHL at the local level, in HDDT design and build, in capturing EBP sources and data, and in data analysis and visualisation theory and best practice.

7.5 Impact Statement

7.5.1 The Evidence Based Prosopographical system

The Evidence Based Prosopographical system proposed by this thesis has the potential to establish a new direction in Digital Humanities research because it brings an information science approach and discipline to the problem of how to digitally represent and study information about Past Human Lives. EBP puts information as Representative Data at the heart of future DH infrastructure. Moving away from a focus on technologies and the digitisation of legacy systems and towards finding, systematising and organising the digital representations of all instances of EBP in Records would enable DH to fully embrace the benefits of the Semantic Web, unite the archive and genealogy, and provide future researchers with infrastructural support at national and (by wider adoption) international levels.

7.5.2 Project Seven

Project Seven has shown that the Independent Researcher model works. Independent Researchers are not significantly disadvantaged compared to group researchers working in 'hubs' or larger national structures. Technologies are rapidly changing and it can be expected that over time, convergence in tool design and use will enable future researchers to benefit more fully from the work of past researchers. Project Seven, and the other projects considered here, have however shown that the greatest complexity and demand on researcher time are in data finding and management. This burden would be significantly eased if the EBP system were to be taken up. If EBP were taken up then the bringing

together of the archive and genealogy would enable greater numbers of Independent Researchers to arise. These new researchers will not have the benefit of academic support unless academia rises to the challenge and builds bridges between the specialist academic researcher and the impassioned non-academic researcher.

Project Seven will be offered to the Quaker academic community as an article in *Quaker Studies* (<https://quakerstudies.openlibhums.org>). It is supported by a GitHub web publication of the Case Study Jupyter Notebooks as a Jupyter Book (<https://jupyterbook.org/en/stable/intro.html>).

Appendix 1 Thomas Hodgkin MD forms the Aborigines Protection Society

'Aborigines Protection Society, first meeting' TRANSCRIPT

WELLCOME LIBRARY FOR THE HISTORY AND UNDERSTANDING OF MEDICINE, DEPARTMENT OF ARCHIVES AND MANUSCRIPTS. HODGKIN FAMILY PAPERS 1996 PP/HO/D, Thomas Hodgkin MD (1798 1866), General Material on Civilisation and Colonialization. Aborigines Protection Society General Materials. PP/HO/D/D148, Minutes of the First Meeting, (3ff), 1837. [Transcribed from cursive script by Kelvin Beer-Jones]

'Aborigines Protection Society, first meeting'

'On Thursday 8th Inst., A meeting on behalf of the Aborigines Protection Society was held at the Friends meeting house, Ratcliff which was both [?] and respectably attended. A little after 7 o'clock, Robert Bell Esq., was called to the chair. The chairman observed that the term aborigines was derived from the Latin, and that in conformity with its derivation, it was applied to the original inhabitants of a country; and he stated that it was the object of this society to assist in protecting and promoting the advancement of defenceless and uncivilised tribes in all parts of the world in which they may exist. The meeting had been convened in this part of London [because?] it was [peculiarly?] situated, in a maritime point of view, and eminently favourable to the wider diffusion of that information which was about to be laid before this meeting. The whole history of the intercourse between the

civilised and uncivilised world was a woeful story and a serious charge against the Government for not having taken adequate pains to collect further information on the subject. It is in some degree to fill up this deficiency that this society has been formed. If we look at the map of the world we can [?] point out a place where some nation or tribe has not become extinct – We ask you to interest yourselves for the presentation of these crimes which have caused much destruction of human life. In Van Dieman's Land the aborigines have been shot like dogs by criminals sent from this country. In S southern Africa within a moderate distance of Natal, upwards of 12,000 Zoolaks have been slain by our fellow subjects. In the islands of the Caribbean Sea the entire aboriginal population has been exterminated. In the United States, whole tribes are being driven from the land of their forefathers and from the soil which they have themselves cultivated; - and we hear of blood hounds conveyed to Florida to hunt down the miserable natives. In fact wherever civilisation has placed its standard, crimes such as these are in perpetration. This requires but to be known to cause one gigantic movement on the part of the British public in support of the object of this society. History has recorded the [serious? ?] to this country during the American war [?] insurrection in Canada, the Indians assisted the loyalists in putting down the rebels. It is an [?] fact that in this 19th century we should be first making the discovery that the coloured men are human beings [?] equal rights to the aborigines and to be assured that they will become an ornament to our system.

Dr Hodgkin moved that an auxiliary society be formed in Ratcliff [? meeting ?] for the purpose of promoting the objects of the Aborigines Protection Society and more especially to collect information from persons recently arrived from abroad and to exert an interest in those who might be going out as colonists or sailors. He observed that though a great amount of valuable information relating to the subjects now under consideration was brought to this part from various parts of the globe very great difficulty had been experienced in collecting it. The formation of the proposed branch society might do much to overcome this difficulty. At present, it is almost by mere accident that the commission of the [grossest?] acts come to light through the accounts of sailors in our merchant vessels. This he had received in statement of a whaler that [?] of vessels engaged in this line would murder the male natives to gain temporary possession of the female who they afterwards abandoned. He would cite this testimony of one who was well known and beloved by many of those to whom they were indebted for the use of this [?] When they were assembled he alluded to the statements which the late Daniel Wheeler had made regarding the profligate conduct of our [?] seamen amongst the islands of the Pacific and of the fearful [?] and destruction of the natives which was the consequence. He then mentioned some facts regarding the injury done to aborigines related in the manuscript journal of James Backhouse another member of the same society. He had visited a spot at which a Kraal of bushmen had settled and had begun to cultivate the ground and make their advances in civilisation; - but from which they had been driven to make way for colonists, and that a

portion of that land had been awarded to a church and received by it. If this can be done by the professors of religion what are we to expect from others? James Backhouse related a fact which formed a striking contrast with this conduct of Europeans. In the hunting parties of the Caffres, the chief men appropriate the best of the [?] until a bushman be present in which case weak and feeble as he is he seen as the [?] because the caffre [recognises ?] in him the oldest and best title to the land and its produce; because the Caffres are taught by tradition that they have spread into the country originally occupied by the bushmen... He had learnt from another indubitable person that Hottentots applying for an allotment of land which they might cultivate for themselves, had been refused lest they should be less willing to devote their services to the colonists.'

END

Appendix 2 Thomas Hodgkin MD's political network

person_id	Name	birth_year	death_year	Target
3386	John Washington	NA	NA	HOD
3371	Jan Tzatzoe	NA	NA	HOD
3366	J H Tredgold	NA	NA	HOD

3359	H B Thorpe	NA	NA	HOD
3357	Perronet Thompson	NA	NA	HOD
3306	Culling Eardley Smith	NA	NA	HOD
3290	Egerton Ryerson	NA	NA	HOD
3244	J F Polk	NA	NA	HOD
3239	x Philip	NA	NA	HOD
3227	Daniel O'Connell	NA	NA	HOD
3219	Standish Motte	NA	NA	HOD
3217	S G Morton	NA	NA	HOD
3201	x Metcalf	NA	NA	HOD
3192	W Martin	NA	NA	HOD
3191	R Montgomery Martin	NA	NA	HOD
3178	x Maconochie	NA	NA	HOD
3172	S Lushington	NA	NA	HOD

3171	C Lushington	NA	NA	HOD
3149	William Kennedy	NA	NA	HOD
3143	Peter Jones	NA	NA	HOD
3128	F Maitland Innes	NA	NA	HOD
3114	Charles Hindley	NA	NA	HOD
3112	R Hill	NA	NA	HOD
3082	R R Gurley	NA	NA	HOD
3061	J J Freeman	NA	NA	HOD
3040	Edward Everett	NA	NA	HOD
3015	Augustus D'Este	NA	NA	HOD
3010	James Davis	NA	NA	HOD
2986	Thomas Clarkson	NA	NA	HOD
2979	Alfred Spencer Churchill	NA	NA	HOD

2972	L A Chamerovzov	NA	NA	HOD
2941	x Blumenbach	NA	NA	HOD
2925	Sax Bannister	NA	NA	HOD
2916	E Baines	NA	NA	HOD
2903	G F Angas	NA	NA	HOD
2884	S J Abington	NA	NA	HOD
2768	Joseph Pease	NA	NA	HOD
2721	Andrew Johnston	NA	NA	HOD
2711	William Howitt	NA	NA	HOD
2710	Luke Howard	NA	NA	HOD
2699	John Hodgkin	NA	NA	HOD
2695	John Herschell	NA	NA	HOD
2684	John Barton Hack	NA	NA	HOD
2680	Samuel Gurney	NA	NA	HOD

2675	Anna Gurney	NA	NA	HOD
2631	James Cropper	NA	NA	HOD
2628	Elliott Cresson	NA	NA	HOD
2607	Edward North Buxton	NA	NA	HOD
2547	R D Alexander	NA	NA	HOD
2535	Robert Howard	NA	NA	HOD
2528	William Forster	NA	NA	HOD
2526	Josiah Forster	NA	NA	HOD
2517	Robert Jun Alsop	NA	NA	HOD
2513	Joseph Sturge	NA	NA	HOD
2486	Frederick Tuckett	NA	NA	HOD
2483	W Thompson	NA	NA	HOD
2365	x Hadfield	NA	NA	HOD
2353	W G Gibson	NA	NA	HOD

2309	Thomas Fowell Buxton	NA	NA	HOD
2280	Jonathan Backhouse	NA	NA	HOD
2115	George Walker	NA	NA	HOD
2114	Edmund Walker	NA	NA	HOD
2108	E T Wakefield	NA	NA	HOD
2024	George C Thompson	NA	NA	HOD
1990	G M Tagore	1,826	1,890	HOD
1972	John James Sturz	1,800	1,877	HOD
1969	x Stuart	NA	NA	HOD
1890	William Smith	NA	NA	HOD
1889	Thomas Southwood Smith	1,788	1,861	HOD
1836	Norton Shaw	NA	1,868	HOD
1744	John Ross	NA	NA	HOD

1732	Frederick W Rogers	NA	NA	HOD
1725	J Robinson	NA	NA	HOD
1713	J Roberts	NA	NA	HOD
1706	Carl Ritter	1,779	1,859	HOD
1687	Anders Retzius	1,796	1,860	HOD
1634	James Cowles Prichard	1,786	1,849	HOD
1448	Roderick Impey Murchison	1,792	1,871	HOD
1359	John Mill	NA	NA	HOD
1207	Joseph Lister	1,827	1,912	HOD
1198	Malcolm Lewin	NA	1,869	HOD
1131	Robert Knox	1,791	1,862	HOD
1117	Richard King	1,811	1,876	HOD
1105	James (1) Kennedy	NA	NA	HOD

1047	J W Jackson	NA	1,872	HOD
967	Thomas (1) Hodgkin	1,798	1,866	HOD
874	Richard Davis Hanson	1,805	1,876	HOD
831	George Grey	1,812	1,898	HOD
830	Charles Edward Grey	1,785	1,865	HOD
802	Richard Thomas Gore	1,799	1,881	HOD
769	x Gawler	1,795	1,869	HOD
755	Francis Galton	1,822	1,911	HOD
730	Robert Nicholas Fowler	1,828	1,891	HOD
713	Joseph Fletcher	NA	1,852	HOD
694	John Fergusson	NA	NA	HOD
678	Edward John Eyre	1,815	1,901	HOD
579	Ernest Dieffenbach	1,811	1,855	HOD
492	John Crawford	1,783	1,868	HOD

414	John Clarke	1,802	1,879	HOD
391	F W Chesson	NA	NA	HOD
346	Robert Campbell	1,829	1,884	HOD
281	x Brougham	1,778	1,868	HOD
265	John Briggs	1,785	1,875	HOD
237	John Bowring	1,792	1,872	HOD
154	William Abraham Bell	1,841	1,921	HOD
81	James (1) Backhouse	1,794	1,869	HOD
28	William (Capt.) Allen	NA	NA	HOD
5	William Adam	NA	NA	HOD
3400	x Wills	NA	NA	HOD
3359	H B Thorpe	NA	NA	HOD
3306	Culling Eardley Smith	NA	NA	HOD
3262	G Ralston	NA	NA	HOD

3244	J F Polk	NA	NA	HOD
3149	William Kennedy	NA	NA	HOD
3082	R R Gurley	NA	NA	HOD
3058	x Foville	NA	NA	HOD
3040	Edward Everett	NA	NA	HOD
2986	Thomas Clarkson	NA	NA	HOD
2938	Samuel Blackburn	NA	NA	HOD
2769	Joseph Whitwell Pease	NA	NA	HOD
2710	Luke Howard	NA	NA	HOD
2699	John Hodgkin	NA	NA	HOD
2680	Samuel Gurney	NA	NA	HOD
2675	Anna Gurney	NA	NA	HOD
2672	Burwood Godlee	NA	NA	HOD
2628	Elliott Cresson	NA	NA	HOD

2611	Edward Carroll	NA	NA	HOD
2593	Joseph B Braithwaite	NA	NA	HOD
2565	J Gurney Barclay	NA	NA	HOD
2535	Robert Howard	NA	NA	HOD
2526	Josiah Forster	NA	NA	HOD
2522	Peter Bedford	NA	NA	HOD
2486	Frederick Tuckett	NA	NA	HOD
2460	Richard Smith	NA	NA	HOD
2438	S Rickman	NA	NA	HOD
2387	Thomas Hughes	NA	NA	HOD
2370	Thomas Harvey	NA	NA	HOD
2309	Thomas Fowell Buxton	NA	NA	HOD
2008	Richard Stephens jnr Taylor	1,843	1,928	HOD

1687	Anders Retzius	1,796	1,860	HOD
1634	James Cowles Prichard	1,786	1,849	HOD
1595	Benjamin Phillips	NA	NA	HOD
1235	John Lubbock	1,834	1,913	HOD
967	Thomas (1) Hodgkin	1,798	1,866	HOD
831	George Grey	1,812	1,898	HOD
830	Charles Edward Grey	1,785	1,865	HOD
713	Joseph Fletcher	NA	1,852	HOD
579	Ernest Dieffenbach	1,811	1,855	HOD
512	Manockjee Cursetjee	1,808	1,887	HOD
465	Frederick Cooper	NA	NA	HOD
403	Henry Christy	1,810	1,865	HOD
81	James (1) Backhouse	1,794	1,869	HOD
28	William (Capt.) Allen	NA	NA	HOD

7	William (1) Adams	NA	NA	HOD
---	-------------------	----	----	-----

Appendix 3 Archival research into the Quaker Committees on the Aborigines

1 Introduction

The Project Seven Report analysis strongly suggests that Quakers played a leading role in the development of the institution of anthropology in Britain from the 1830s until the late 1860s, and this was shown by the P7 report to be largely through the Quaker networks of Thomas Hodgkin MD (1790-1866), who began his institution building first with the Quaker Committee on the Aborigines, then the Aborigines Protection Society and finally with the Ethnological Society of London. The claims made by the P7 Report are strengthened by detailed analysis of manuscripts held in Quaker archives (Friend's House Archives, London), and this appendix is a report of those manuscript researches. It supports the Project Seven Report, and its claims, in Appendix 7.

The London Doctor of Medicine and Quaker Thomas Hodgkin MD (1798–1866) had a deep interest in the scientific study of Man, which he developed in part through his contacts with the Quaker James Cowles Prichard (1786–1848). By 1835, Prichard had already established himself in Bristol as a medical practitioner and the leading authority in Britain for the scientific study of Man.

To develop his own interest in the scientific study of Man, Hodgkin took advantage of a philanthropic and evangelical Quaker concern for the plight of aborigines that emerged very early in the new century and came to a head in 1837 (as the following sections on the Minute Books of the Quaker London Yearly Meeting and the Minute Books of the Quaker Meeting for Sufferings will show).

In 1837, utilising Pritchard's methods of information gathering which he developed and practised at Bristol, Hodgkin began to build in London institutions dedicated to the systematic collection of information about the condition of aborigines in the British Settlements. He collected this information initially from the dockland maritime communities in London at Ratcliffe (a locality in the London Borough of Tower Hamlets).

It is noteworthy that Charles Darwin would soon after assist both Hodgkin and Pritchard in designing and promoting similar information gathering processes that they then collectively continued to deploy among Britain's maritime communities for the next twenty or so years. It is in the scientific, fact collecting and testimony based information gathering activities of Pritchard and Hodgkin at British ports, which at the time were trafficking widely with the colonial territories, that the beginnings of the science of anthropology in Britain lie.

Pritchard and Hodgkin were both Quakers and it is Quakers in Britain who in the early nineteenth century unwittingly gave their support to and supplied the resources which would nurture these first signs of anthropology.

The Quaker London Yearly Meeting and its Meeting for Sufferings struggled during the 1830s and 1840s with repeated news of atrocities performed on aborigines by settlers throughout the British Settlements, when trying to provide spiritual guidance to the increasing numbers of Quakers emigrating to the Settlements and thereby interacting often disastrously with aborigines, and with conflicts in North Carolina and later Upper Canada stretching the ability of Quakers in Britain to make meaningful interventions in the Settlements that balanced all of their often conflicting concerns.

Nonetheless, Quakers throughout the 1830s and 1840s actively supported the endeavours of Hodgkin in institution building, for example when on 5 June 1837 Quakers (including Thomas Hodgkin's younger brother John Hodgkin Jnr) formed the Committee on the Aborigines⁵⁸¹ and then three days later Hodgkin formed the Aborigines Protection Society (APS) at Ratcliffe Quaker Meeting House.

In the 1840s Hodgkin used the APS as a springboard from which to bring Pritchard to London from Bristol. After some time wrangling with some of the newly formed scientific institutes, Pritchard, with the enthusiastic support of Hodgkin, became the founding chair of the Ethnological Society of London in 1843. In 1871 the Ethnological Society of London would become the Royal Anthropological Institute.

⁵⁸¹ It will be termed variously in the primary sources, but we chose to use the form most commonly employed at the time.

2 Description of the Quaker archives

I began my search with the minutes of London Yearly Meeting 1830–1850; this is the record of the annual Quaker gathering at national level, and it is the highest governing body for Quakers in Britain. I found seventeen entries referring to aborigines (or other derivative terms). Many of these references mentioned the standing committee of the Yearly Meeting, called the Meeting for Sufferings. Here I found eighteen entries. Meeting for Sufferings entries sometimes mentioned London Quarterly Meeting and other Monthly (local) Quaker meetings in the London area. London Quarterly Meeting offered up six entries. When at the Wellcome Institute, I found the notes of the first meeting of the Aborigines Protection Society which had been held at Ratcliffe Meeting House, and so I then examined the records of that local Meeting too. This resulted in the discovery of a further nine entries. Finally, I examined the local meeting records for the Devonshire House Meeting attended regularly by Thomas Hodgkin and found six entries there. In all fifty-six entries mentioning aborigines or derivatives were found.

Quaker corporate records follow a traditional format of mostly routine and standardised reports made about the life of the Meeting each year. Therefore, the minute book records create an easy pattern to follow. Very few concerns of the day that are outside of the routine life of the community make their way to the minutes of London Yearly Meeting, and when they do they stand out. It is significant that so many entries referring to aborigines

were found in this period and not in the decade or so before, or after. In most years, the matter of the plight of aborigines was the only non-routine item recorded in the minute books of London Yearly Meeting 1831–1846.

My review of Quaker corporate records began with the Minutes of London Yearly Meeting in 1820, and the first record citing ‘heathen’ in a minute calling for the introduction of the Gospel to aborigines dates from 1831.⁵⁸² In 1832, this concern was repeated at London Yearly Meeting⁵⁸³ with the decision that an exploratory committee be established. This is also the first entry where a list of names appears. At fifty-two members it is unusually large for a Quaker committee, and whether this considerable number indicates a growing interest in evangelicalism or in philanthropy is not clear. For those taking a philanthropic lead, the conversion of aborigines to Christianity was often seen as a first step towards their

⁵⁸² ‘The consideration whether the time is come when our society is called upon to take a more decided part as a body than it has hitherto done in communicating to the heathen the knowledge of the Gospel has again occupied the deep and solid attention of this meeting and under a renewed feeling of the great importance of the subject it is referred to the attention of the next yearly meeting’ (Library of the Society of Friends 1831, 5th Month 24, 46).

The evangelical movement in Britain and its relationship to emergent modernity in all its forms are explored in (Stubenrauch 2016).

⁵⁸³ ‘This meeting ... under a solemn sense of the importance concludes to refer further consideration of it [the subject] to a committee of this meeting which is to make such a report thereon at it may feel proper to the next yearly meeting. The Committee is as follows: Sylvanus Fox, Josiah Forster, George Jones, Richard Cockin, John Talwin Shewell, Joseph Marriage, Joseph John Gurney, Thomas Shillitoe, Cornelius Hanbury, John Candler, Samuel Tuke, Thomas Maw, Samuel Lloyd, George Stacey, John Barclay, Richard T Cadbury, Joseph Foster, John Young, William Allen, Edward Pease, George Crosfield, Richard Ball, William Ball, Peter Bedford, Luke Howard, John Brown, William Dillworth Crewdson, William Foster Reynolds, William Binns, Isaac Braithwaite, Robert Forster, Robert Jowitt, John Capper, Wyatt George Gibson, Joseph Jewell, James Ireland Wright, Richard Dykes Alexander, William Gundry, James Forbes, Richard Barrett, Sheldon Dudley, John Rickman, Edward Smith, Samuel Capper, John Dymond, Thomas Bigg, Robert Were Fox, William Boulton, Barnard Dickinson, John Talbot, Isaac Wilson, Dykes Alexander, Joseph Storrs Fry, Jonathan Hutchinson, Henry Newman’ (Library of the Society of Friends 1832, 5th Month 31, 212-214).

improvement and emancipation. The list includes Joseph Gurney, but not members of Thomas Hodgkin's family or that of Joseph Pease (the only Quaker MP). Gurney may well be included here not because of his family connection to Sir Thomas Fowell Buxton, but rather because of his evangelical leanings (Hamm 2013, 71).

In 1833, this unusually large Quaker 'Committee of Enquiry' reported back to London Yearly Meeting that it was unable to propose a way forward with the concern that was its remit.⁵⁸⁴ Instead the Committee of Enquiry made a plea for the plight of aborigines ('heathens'), which it had uncovered during its enquiry.⁵⁸⁵ It seems clear that while the Committee of Enquiry began with an evangelical concern, what it produced was an emerging philanthropic concern. This it may have happened because of testimony received as a result of opening up Quaker lines of communication about aborigines with Quaker colonists.⁵⁸⁶

3 Table of manuscripts

⁵⁸⁴ 'On considering the important subject referred to us, we have been led into a deep feeling of interest therein; but after much deliberation and the free interchange of sentiment, we have come to the conclusion that we cannot recommend to the Society as a body the adoption of any specific measure at the present time' (Library of the Society of Friends 1833, 6th Month 5, 394).

⁵⁸⁵ 'The deplorable condition of the heathen and the degrading circumstances under which they are living have been felt at this time, as well as in former years to be truly affecting. And although no way appears to open for the Society to adopt any specific measure in order to communicate to them the knowledge of the truths of the Gospel, we earnestly recommend their benighted condition to the Christian sympathy and frequent remembrance of all our members' (Library of the Society of Friends 1833, 6th Month 5, 398).

⁵⁸⁶ How colonial networks function, especially as conduits with London elites, is fully analysed in (Laidlaw 2005b), see especially p. 4.

Search parameters: Any recorded interest in ‘aborigines’, ‘natives’ and other derivatives of the term, 1830–1850.⁵⁸⁷ Search performed 6 September 2017–19 December 2017.⁵⁸⁸

Year	London Yearly Meeting	Meeting for Sufferings	Manuscript Reference ⁵⁸⁹
	Volume 44		
1831	1831 5th Month 24 (p. 46)		
1832	1832 5th Month 31 (pp. 212–214)		
1833	1833 6th Month, 5 (pp. 394–398)		
	1833 Testimony Extract		

⁵⁸⁷ It would be good to look back to earlier years in detail; however, a casual survey did not reveal matters relating to aborigines 1820–1830.

⁵⁸⁸ Following the leads from manuscript to manuscript in my desire to make a comprehensive search resulted in searching the records of several Quaker groups. My cataloguing is incomplete, 6 entries at ‘Devonshire House Monthly Meeting’, 9 entries at ‘Ratcliff Monthly Meeting’, 6 entries at ‘London Quarterly Meeting’; and 5 entries from ‘Meeting for Sufferings’ are to be re-examined. How Quaker Meetings in Britain are constituted and organised both today and historically are set out in the Fifth Edition of *Quaker faith & practice*, Chapters 4–9. <http://qfp.quaker.org.uk/chapter/4> (accessed 12/5/2023)

⁵⁸⁹ Photographs of most entries were made and have been marked for reference.

1834	1834 5th Month, 30 (p. 448)		
	1834 5th Month, 30 (p. 479)		
1835	1835 5th Month, 26 (p. 523)		
1836	No entries	No entries	
	Volume 45		
1837	1837 5th Month, 31 (p. 108)	Volume 44	
		1837 5th of the 6th Month (p. 394)	
		1837 7th of the 7th Month (pp. 419– 420)	

		1837 4th of the 8th Month (p. 429)	
		1837 7th of the 9th Month (p. 539)	
1838	1838 5th Month, 29 (pp. 155–156)		
		1838 5th of the 10th Month (p. 542)	
1839	1839 5th Month, 25 (pp. 216–217)		
		1839 7th of the 6th Month (pp. 587 and 589)	
		1839 1st of the 11th Month (p. 612)	

		1839 6th of the 12th Month (p. 616)	
		1839 Memorial to Lord John Russell (pp. 623–625)	
		Volume 45	
1840		1840 7th of the 2nd Month (p. 5)	
	1840 5th Month, 26 (p. 276)		
		1840 5th of the 6th Month (pp. 36–37)	
		1840 7th of the 8th Month (p. 48)	
		1840 6th of the 11th Month (p. 72)	

	1840 (pp. 293–294)		
1841	1841 5th Month, 25 (p. 334)		
	1841 5th Month, 25 (p. 342)		
1842		1842 7th of the 10th Month (pp. 195– 197)	
1843	No entries	No entries	
1844		1844 3rd of the 5th Month (p. 286)	
	1844 5th Month, 30 (pp. 499–500)		
1845		1845 16th of the 5th Month (pp. 345– 346)	

	1845 5th Month, 22 (pp. 541–542)		
1846		1846 16th of the 5th Month (p. 439)	
	1846 5th Month, 26 (p. 606)		
1847		1847 6th of the 8th Month (pp. 533– 534)	

4 The Minute Books of London Yearly Meeting and the Minute Books of the Meeting for Sufferings (1831–1846)

Quakers in Britain's interest in aborigines at a national level is recorded in the minutes of the Quaker London Yearly Meeting for most years from 1831 to 1846⁵⁹⁰ (with only a few exceptions: 1836 and 1842–1844)⁵⁹¹ and in the national standing committee of London

⁵⁹⁰ At which time, with the formation of the Ethnological Society of Britain complete in 1843, our interest ceases.

⁵⁹¹ Interestingly, RAI A111 Papers of the Aborigines Protection Society notes 'missing' data in similar years, 'The annual reports for the following years are missing: 1841–6'. I will try to ascertain why these are dearth years for both the Quaker Committee and the APS as far as activity/reporting goes.

Yearly Meeting titled Meeting for Sufferings. Through the minutes of these two groups we can trace the emergence of a concern for the promulgation of the Gospel among aborigines, which would rapidly transform into a philanthropic concern for the plight of aborigines.

After the hiatus of the publication of the House of Commons Select Committee Report on Aborigines in 1837 (Papers 1839) the interest of British Quakers at national level wanes, but it was then to be replaced by the more vigorous involvement of both Pritchard and Hodgkin, working among the scientific clubs then active in London to promote the scientific study of Man.

We can see in a reading of the minutes of London Yearly Meeting (LYM) and the Meeting for Sufferings (MfS) for this period the steps taken by Quakers in Britain at a national level to address both Gospel and philanthropic matters relating to aborigines, and how Hodgkin gets involved in the emerging Quaker philanthropic concern. He does this initially through his father and his role within the Quakers at national level, but he quickly goes on to use this Quaker philanthropic concern for the plight of aborigines to build a platform from which he can pursue his and Pritchard's closely related scientific concern, to establish institution building in anthropology within London's wider intellectual society.

The Minutes of London Yearly Meeting in 1831 record a concern for the aborigines solely in terms of promulgation of the Gospel:

The consideration whether the time is come when our society is called upon to take a more decided part as a body than it has hitherto done in communicating to the heathen the knowledge of the

Gospel has again occupied the deep and solid attention of this meeting and under a renewed feeling of the great importance of the subject it is referred to the attention of the next yearly meeting.

(Library of the Society of Friends 1831, 5th Month 24, 46)

In 1832⁵⁹² this concern is repeated at London Yearly Meeting with the important addition that an exploratory committee be established:

This meeting ... under a solemn sense of the importance concludes to refer further consideration of it [the subject] to a committee of this meeting which is to make such a report thereon at it may feel proper to the next yearly meeting. The Committee is as follows:

Sylvanus Fox	Edward Pease	William Gundry
Josiah Forster	George Crosfield	James Forbes
George Jones	Richard Ball	Richard Barrett
Richard Cockin	William Ball	Sheldon Dudley
John Talwin Shewell	Peter Bedford	John Rickman
Joseph Marriage	Luke Howard	Edward Smith
Joseph John Gurney	John Brown	Samuel Capper
Thomas Shillitoe	William Dillworth Crewdson	John Dymond
Cornelius Hanbury	William Foster Reynolds	Thomas Bigg
John Candler	William Binns	Robert Were Fox
Samuel Tuke	Isaac Braithwaite	William Boulton
Thomas Maw	Robert Forster	Barnard Dickinson
Samuel Lloyd	Robert Jowitt	John Talbot
George Stacey	John Capper	Isaac Wilson
John Barclay	Wyatt George Gibson	Dykes Alexander
Richard T Cadbury	Joseph Jewell	Joseph Storrs Fry

⁵⁹² This Yearly Meeting opens with a concern: 'We acknowledge our reverent thankfulness to the Preserver of men, that the pestilence which has visited several parts of the kingdom, since we last met, is now very much diminished. The ravages of this disease have been far greater in other nations than ours...' It might be helpful to find out what this epidemic was and its extent.

Joseph Foster
John Young
William Allen

James Ireland Wright
Richard Dykes Alexander

Jonathan Hutchinson
Henry Newman⁵⁹³

(Library of the Society of Friends 1832, 5th Month 31, 212-214)

In 1833 the 'committee appointed to consider communicating to the heathen the knowledge of the Gospel'⁵⁹⁴ reported back to London Yearly Meeting. While the committee and the Meeting did not find a way then to further their interest in the promulgation of the Gospel, a new concern for the plight of aborigines is made by the committee and is recorded with some power by the London Yearly Meeting:

On considering the important subject referred to us, we have been led into a deep feeling of interest therein; but after much deliberation and the free interchange of sentiment, we have come to the conclusion that we cannot recommend to the Society as a body the adoption of any specific measure at the present time.

The deplorable condition of the heathen and the degrading circumstances under which they are living have been felt at this time, as well as in former years⁵⁹⁵ to be truly affecting. And although no way appears to open for the Society to adopt any specific measure in order to communicate to them the knowledge of the truths of the Gospel, we earnestly recommend their benighted condition to the

⁵⁹³ These 55 names are written in the minute book in line, but listed out here for ease of review and they are the seed community that Hodgkin will use in his institution building.

⁵⁹⁴ The committee on the Gospel may be thought of as a precursor to the Committee on the Aborigines.

⁵⁹⁵ This suggests that more evidence may be found in the Minutes of earlier Yearly Meetings but this research did not have the scope to look for it.

Christian sympathy and frequent remembrance of all our members.’ (Library of the Society of Friends 1833, 6th Month 5, 394-398).

The next year, on 27 May 1834, London Yearly Meeting considered an Epistle from North Carolina Meeting urging that North Carolina Meeting (which was now under the care of London Yearly Meeting) be allowed to evict ‘coloured peoples’⁵⁹⁶ in that territory in return for ‘some pecuniary assistance’. It is clear that this assistance would be a compensation payment to coloured peoples for displacement.⁵⁹⁷

Given the level of awareness of concern for the plight of aborigines growing among London Yearly Meeting based on the reports of the ‘committee appointed to consider communicating to the heathen the knowledge of the Gospel’ via the Meeting for Sufferings of the previous year, this Minute from North Carolina proved difficult to accept without further consideration and it was taken away for that purpose. The matter was not resolved until the following year. Here we see Quakers in Britain conflicted between the needs of Quaker settlers in the colonial territories on the one hand and the plight of aborigines with whom those settlers come into contact on the other:

An epistle has now been received and read from the Meeting for Sufferings for North Carolina held 14th of the 4th month, 1834 on the subject of the coloured population under the care of our Friends there, it is concluded to propose to the yearly meeting that liberty should be allowed to this meeting

⁵⁹⁶ The term ‘coloured people’ is here taken to refer to aborigines.

⁵⁹⁷ The background of the unrest in North Carolina at this time is unclear in the sources and needs further research, this is not within the scope of this research.

to render some pecuniary assistance on the occasion, with an especial view to the removal, with their own consent, of the said people of the State of North Carolina.⁵⁹⁸ (Library of the Society of Friends 1834, 5th Month 30, 448)

London Yearly Meeting then adjourned and probably revisited the matter three days later on 30 May, if we can infer that this new minute is in response to the request from North Carolina:

The following minute has been brought in and read and this meeting recommend the subject to which it refers to the close attention of the Meeting for Sufferings with directions to make report thereon to the yearly meeting next year.

Committee on Epistles 29th May 1834, This committee has been introduced into much concern on behalf of the aborigines of North America and suggests to the Yearly Meeting to recommend the subject to the close attention of the Meeting for Sufferings with directions to make report thereon.

Samuel Gurney

Clerk. (Library of the Society of Friends 1834, 5th Month 30, 479)

The following year (1835) the North Carolina matter of the displacement of natives seemed to have been resolved at London Yearly Meeting and one can imagine probably not to the pleasure of Hodgkin:

This meeting approves of the proceedings of the Meeting for Sufferings in regard to the coloured people under the care of the yearly meeting of North Carolina.⁵⁹⁹

⁵⁹⁸ No further mention is made of this request and it is not clear to me from the style of this record whether it can be taken as granted.

⁵⁹⁹ This matter would be addressed and resolved by the Society in 1840.

Samuel Gurney, clerk. (Library of the Society of Friends 1835, 5th Month 26, 523)

In 1836 London Yearly Meeting did not minute matters relating to aborigines, but in 1837 the Meeting for Sufferings appointed a Committee on the Aborigines⁶⁰⁰ which was to be a subcommittee of the General Committee (of Meeting for Sufferings):

The following friends are appointed to take such steps as they may be enabled, to further the objects of the Yearly Meeting, as conveyed in its minute on the subject of the present state of the aborigines of the British Colonial Possessions, particularly as it respects those of the Indians of Upper Canada, viz William Allen, Thomas Christy, John Hodgkin Jun, John Thomas Barry, William Forster, Robert Forster, Abram Rawlinson Barclay, George Stacey, Robert Howard and Henry Knight Jun.⁶⁰¹ (Library of the Society of Friends 1837b, 6th Month 5, 394).^{602 603}

The Committee on the Aborigines then made an immediate report to London Yearly Meeting through the General Committee:

The following minute from the General Committee has been brought in and read. This meeting adopts the suggestion therein contained and desires the Meeting for Sufferings to pay close attention to the subject and to act in it at their discretion. (General Committee 6th Mo 1st 1837)

⁶⁰⁰ In the primary sources the name of this Quaker committee varies.

⁶⁰¹ Robert Bell would chair the first meeting of the Aborigines Protection Society, but he is not a member of the Committee of the Aborigines made here by the Meeting for Sufferings and only William Allen, Robert Forster and George Stacey were also members of the previous year's 'committee appointed to consider communicating to the heathen the knowledge of the Gospel'.

⁶⁰² This is the first mention of a Quaker Committee, and among its members is Thomas Hodgkin's father.

⁶⁰³ The first meeting of the Aborigines Protection Society would take place three days later on Thursday 8 June at Ratcliffe Meeting House.

This committee having under its serious consideration the circumstances of the aborigines of the British Colonial possessions, particularly the Indians of upper Canada submits to the yearly meeting the propriety of recommending the close attention of the subject to the Meeting for Sufferings.

Young Sturge

Clerk (Library of the Society of Friends 1837a, 5th Month 31, 108)

In 1837 the House of Commons Aborigines Select Committee report was published and Hodgkin, together with Sir Thomas Fowell Buxton MP, the Select Committee chair, formed the Aborigines Protection Society (APS) at Ratcliffe Quaker Meeting House in London, three days after the establishment of the Quaker Committee on the Aborigines. Hodgkin acts as secretary to the newly formed APS (Appendix 1). His father John Hodgkin Jnr sits on the Quaker Committee on the Aborigines.

The Committee on the Aborigines begins to meet monthly and makes regular reports back to the Meeting for Sufferings:

The friends appointed on the subject of the present state of the aborigines of the British Colonial Possessions are continued: the following friends are now added. Josiah Forster, William Hargreave, Samuel Darton, John Sanderson, Peter Bedford, Thomas Norton Jun, Richard Barrett, John Hamilton, Edwd Harris, Robert Alsop Jun, John Kitching, Joseph Storrs, Geo Holmes, Joseph Shewell, John Bell, Joseph Neatby, John Barclay, and Joseph Talwin Foster. (Library of the Society of Friends 1837b, 7th Month 7, 419-420).

Josiah Forster reports that the committee on the state of the aborigines in the British Colonies met and attended to their appointments: the committee is continued. (Library of the Society of Friends 1837b, 8th Month 4, 429).

The committee on the state of the aborigines in the British Colonies is continued. (Library of the Society of Friends 1837b, 9th Month 7, 539).

The relationship between the Aborigines Protection Society and the Quaker Committee on the Aborigines is unclear and indeed, given the common membership of both organisations, this lack of clarity might have been manifest at the time and perhaps even deliberate (at least in the mind of Hodgkin).

Over the next two years with the publication of several reports in the public domain, and a memorial to the Secretary of State, the cause of aborigines is taken up once again by the Quaker Committee on the Aborigines, as reported at London Yearly Meeting⁶⁰⁴ on 30 May 1838:

The remaining portion of the selected minutes from the Meeting for sufferings has been read, amongst these minutes is a memorial to the Earl of Durham on the subject of the Indians in Canada, also a report from the committee of Baltimore Yearly Meeting on Indian affairs,⁶⁰⁵ both of which documents this meeting directs to be printed and circulated under the care and at the discretion of the Meeting for Sufferings (Sufferings 1839), together with any other information on these subjects which the said meeting may think desirable (Sufferings 1838).

⁶⁰⁴ The similarities in the publications of both the APS and the Quaker Committee and their shared membership are such that we can regard them as in effect one and the same.

⁶⁰⁵ I have to hand most of the publications made by both organisations.

Some information on the state of the aborigines of the British Colonial possessions has been laid before this meeting by the Meeting for Sufferings which is desired to bring to the next yearly meeting such further information as it may receive on this subject.

This meeting has been deeply interested with some information now communicated respecting the circumstances of the African race, both in their native land and in the colonies of this country; and understanding that much pecuniary assistance will be required to promote their welfare as respects their education in our colonies and otherwise this meeting thinks it right to encourage a very liberal subscription among our members throughout the country to be applied under the direction and at the discretion of the Meeting for Sufferings in connection with the object above mentioned. (Library of the Society of Friends 1838a, 5th Month 29, 155-156)

The Committee on the Aborigines has, it seems, been very active, as is evidenced by the level of intelligence that it is now regularly feeding into London Yearly Meeting and the Meeting for Sufferings:

Robert Forster reports that the committee appointed on the subject of the Aborigines have met since last month, the committee continues. (Library of the Society of Friends 1838b, 10th Month 5, 542).

On 29 May 1839, the minutes of London Yearly Meeting record a continuing interest in the welfare of aborigines but do not propose any further measures now, other than that Meeting for Sufferings is 'encouraged to continue its attention to the subject and to transmit to our next yearly meeting any further information which it may obtain'.

Perhaps the most important entry for us in the record of the Meeting for Sufferings comes in June 1839 when we learn that Thomas Hodgkin has joined the Committee on the Aborigines:

The list of correspondents has now been called over, and the following alterations and additions have been proposed and agreed to viz;

In London,⁶⁰⁶ Thomas Hodgkin added. (Library of the Society of Friends 1839, 6th Month 7, 587-588).

It may well be that Thomas Hodgkin has been brought into the Committee on the Aborigines to help work on the response to information received about the 'Indians of Canada whom the policy and treatment of the late Lieutenant Governor, Sir F B Head are still depriving of their land and sending into the Western Forests' and to assist in the drafting of the Memorial to Lord John Russell. In addition the committee continues reporting on the important theme of the behaviour of colonists:

A minute has been brought in from the committee on the Aborigines for the adoption of this meeting which has been considered, and is, with a small alteration agreed to, the minute is as follows, copies of the same are to be forwarded to the different Quarterly Meetings.

The case of members of our society who may be contemplating emigration to distant colonies has been brought under the notice of this meeting; and we think it right to remind our dear friends who

⁶⁰⁶ Branch Meetings took place outside of London in Bristol, Liverpool and perhaps Edinburgh. No records survive.

may be so circumstanced how much the steps they take make affect not merely the interests of humanity but moral and Christian principle.

We would entreat those who may establish themselves in newly settled countries to reflect upon the responsibility which attaches to them when they are the neighbours of uncivilised and heathen tribes. It is an awful but indisputable fact that most settlements of this description, besides dispossessing the natives of their lands without equivalent, have hitherto been productive of incalculable injury to the moral and physical condition of the native races which have been thereby more or less reduced in numbers, and in some instances completely exterminated. Earnestly therefore do we desire that all those under our name who may emigrate to such settlements may be careful neither directly nor indirectly to inflict injury upon the natives, but that they may on the contrary in their whole conduct exhibit the practical character of that religion that breathes 'peace on earth and goodwill toward men.' As this is their aim they will not only exert themselves to check the evils which are but too generally inflicted by the whites upon their feebler neighbours but will be solicitous to do their part in endeavouring to diffuse amongst them the blessings of civilisation and Christianity, which will prove the best means of preventing their extermination, and of raising them to the full enjoyment of their rights. (Library of the Society of Friends 1839, 11th Month, 1, 612).

At the end of 1839, the first attempt of the Committee on the Aborigines to directly influence government policy emerges in a Memorial to Lord John Russell, Secretary of State for the Colonies:

A Memorial has been produced and read from the committee on the Aborigines addressed to Lord John Russell as secretary of state for the colonies, which the said committee suggests for the adoption of this meeting, the memorial has been several times read and considered, and is with a few alterations been adopted and signed by the clerk on behalf of the meeting, the case of presenting the

same is left to the following friends, viz. William Allen, Samuel Gurney, Thomas Hodgkin, Robert Forster, George Stacey and John Hodgkin Jun. (Library of the Society of Friends 1839, 12th Month 6, 616).

In February of the following year, 1840, it is noted in the records of the Meeting for Sufferings that 'Samuel Gurney reports that the Memorial to Lord John Russell on the subject of the aborigines was presented to him on the 8th of last month' (Library of the Society of Friends 1840b, 2nd Month 7, 5).

In 1840 the Meeting for Sufferings reported back as requested the previous year to the London Yearly Meeting, which now records:

In pursuance of the recommendation of the meeting last year the Meeting for Sufferings has produced some further interesting information on the subject of the aborigines which has been now read - the said meeting is desired to print and circulate such portion of this information as it may think desirable. (Library of the Society of Friends 1840a, 5th Month 26, 276)

A few days later the London Yearly Meeting minutes record their adoption of the Meeting for Sufferings minute of 1 November 1839:

We would entreat those who would establish themselves in newly settled countries to reflect to the responsibility which attaches to them when they are the neighbours of uncivilised and heathen tribes. It is an awful but indisputable fact that most settlements of this description, besides dispossessing natives of their land without equivalent, have hitherto been productive of incalculable injury to the moral and physical condition of the native races, which have been thereby more or less reduced in

numbers and in some instances completely exterminated. Earnestly therefore do we desire that all those under our name who may emigrate to such settlements may be careful neither directly or indirectly to inflict injury upon the natives; but that they may on the contrary, in their whole conduct, exhibit the practical character of that religion which breathes, 'Glory to God in the highest on earth peace goodwill toward men.' As this is their aim, they will not only exert themselves to check the evils which are but too generally inflicted by the whites upon their feebler neighbours, but will be solicitous to do their part in endeavouring to diffuse amongst them the blessings of civilisation and Christianity, which will prove the best means of preventing their extermination, and to the raising of them to the full enjoyment of their rights.' (Library of the Society of Friends 1840a, 5th Month 26, 293-294)

This minute possibly finally addresses the concern not fully dealt with some six years earlier in 1835, after a request from the Meeting for Sufferings of North Carolina in the previous year's Epistle, seeking permission to raise funds with which to evict aborigines (coloured peoples) from the province, was not addressed (see above).

During the remainder of 1840, the Committee on the Aborigines continued to meet regularly and circulate documents. In August, the committee reorganised:

This list of the aborigines committee has been revised and it is agreed that it now consist of the following friends, William Allen, John Hodgkin Jun, John Thomas Barry, William Forster, Robert Forster, George Stacey, Henry Knight, Josiah Forster, John Sanderson, Peter Bedford, Thomas Norton

Jun, Edward Harris, Robert Alsop Jun, Joseph Storrs, Joseph Talvin Foster, William Grimshaw Jun, Edward Paull and Thomas Hodgkin.⁶⁰⁷ (Library of the Society of Friends 1840b, 8th Month 7, 48).

The committee on the state of the aborigines have brought in a document containing Christian Counsel to emigrants and evincing an interest on behalf of the aborigines which has been read and the concern cordially united with by this meeting. The committee are encouraged to give the subject further consideration and present the document at our next meeting for its adoption. (Library of the Society of Friends 1840b, 11th month 6, 72).

With the newly formed Committee on the Aborigines there seems to be a change of pace, and a lack of certainty over the committee's remit or effectiveness among the members will soon emerge. We begin to see a slipping away of recent heightened concern for aborigines at national level among Quakers in Britain.

In 1841 James Backhouse gave an account of his recent travels in Australia, Van Diemen's Land, Mauritius and South Africa accompanied by George Washington Walker. An appeal was also made to Monthly Meeting in the UK that they correspond with and seek to maintain an interest in the welfare of friends who emigrate.

In 1842 doubt began to emerge about the continuance of the Committee on the Aborigines:

The following report was received from the committee on the aborigines.

At a committee on the aborigines held on 7th of the 10th month 1842.

⁶⁰⁷ Thirteen members retiring and only William Grimshaw Jnr and Edward Paull joining, perhaps signalling that the committee is in difficulty.

The committee of the Meeting for Sufferings on the subject of the aborigines in considering (before the occurrence of the last yearly meeting) the propriety of presenting a report seriously entertained the question whether it might not be right to propose the discontinuance of the committee altogether, but remembering the interest which had been from time to time expressed in the yearly meeting on the subject, and observing in a retrospect of the committee's labours in previous years that several subjects of practical importance had come before them and also before the Meeting for sufferings in connection with the wrongs of the aborigines, they are reluctant to take this course.

They have met since the yearly meeting and have again looked both at the general question and also at some of the branches of it which they apprehend that a body acting on behalf of the society could most advantageously entertain.

At the same time they are desirous that the Meeting for Sufferings should devote a little time to the serious consideration of the whole subject; and decide for itself what is most advisable to be done in the interval between the present time and the occurrence of the next yearly meeting; and that if any committee be continued the meeting would revise the present list.

They also submit to the Meeting for Sufferings that if a committee be continued that all papers received by the meeting bearing on the interests of the aborigines should be referred to such committee for attention, and that to it should be the examination and printing of the materials of this description which were presented to the last yearly meeting.

A copy of this minute to be taken to the Meeting for Sufferings.

This meeting agrees to the suggestion in the report and desires to encourage the committee to continue their attention to the subject under their care in relation to their appointment. James Bowden, William Nash, and Joseph Sturge are now added to the committee. (Library of the Society of Friends 1842, 10th Month 7, 195-197).

The appeal that Monthly Meetings in the UK take a pastoral interest in those of their former members who have emigrated was repeated in 1844 when the Meeting for Sufferings makes a request:

The committee on the state of the aborigines are requested to present a brief report of their proceedings during the past year to the approaching yearly meeting. (Library of the Society of Friends 1844, 5th Month 3, 286).

In 1845 the work of the Committee on the Aborigines continued, now with Thomas Hodgkin clearly taking the lead, in perhaps a last attempt to keep the Quaker committee alive. He reported to the Meeting for Sufferings:

The following report has been presented by the Aborigines Committee.

The Aborigines Committee of the Meeting for Sufferings has with scarcely any interruption met monthly during the past year in the course of which it has received information respecting the aborigines in several of the British Colonies and also from parts which are independent of the country.

With little exception these accounts have exhibited the feeble, the ignorant and the pagan suffering in various ways from the oppressive hands of their civilised fellow creatures who abusing the superiority conferred by the knowledge and resources which it is their privilege to have received, dishonour the Christian name which they unworthily profess. Amongst these cases may be mentioned numerous encroachments on the lands and other possessions of North American Indians – the kidnapping of Indians of Guyana by Brazilian slave dealers – the deportation of negroes and coolies under the designation of free emigrants – the destruction of Australian natives by means of poison introduced into their food, which is perpetrated with impunity not from the inhumanity or indifference of the officers of government but from the defect of the law which gives no validity to the testimony of

aborigines, and the efforts which are being made to obtain possession of land in New Zealand in violation of the rights of the native population.

Although these circumstances are far from evincing any improvement in the state of things in relation to the subject, the committee has not seen its way to take any particular steps, except that very recently individual members of the committee have used some exertion in behalf of the right of the natives of New Zealand.

The committee has published a small pamphlet consisting of the two reports of the preceding years, with an appendix containing information respecting the Indians of British North America, as well as those tribes more immediately under the notice of the American yearly Meeting.

During the past year the committee has been made [gen---vedly] sensible that whilst there exists numerous cases demanding sympathy, coming almost within the range of personal observation, it is extremely difficult to sustain, or even to excite an interest in behalf of those distant branches of the human family for whose sake this committee has been appointed.

The committee is apprehensive that the pamphlets relating to the aborigines already published by direction of the Meeting for sufferings are far from being generally diffused amongst the members of the Society, and that even the libraries of meetings are not completely supplied with them.

In conclusion the committee would again recommend the subject for the increased attention and sympathy of Friends.

Signed on behalf of the committee.

Devonshire House 14th of the 5th Month 1845.

Thomas Hodgkin. (Library of the Society of Friends 1845, 5th Month 16, 345-346).

Hodgkin's attempts were ultimately unsuccessful, because on 5 May 1846 London Yearly Meeting minutes recorded:

A report has now been brought in from the Meeting for Sufferings as to the amount received from the different Quarterly Meetings in conformity with the minute of last year recommending a subscription on behalf of the coloured population in the West India Islands and elsewhere and the aborigines of different countries. The said report contains also information as to the appropriation of a part of the said fund. It is satisfactory to this meeting which desires the Meeting for Sufferings to transmit annual to this meeting the state of the fund. (Library of the Society of Friends 1846a, 5th Month 26, 606)

On 16 May the Committee on the Aborigines made its last report to the Meeting for Sufferings with this lament:

Nothing has occurred since the last yearly meeting to [warrant?] the belief that any important changes have been made to advance the moral and physical improvement of the aborigines in the British colonies, or that they are enjoying to the full extent what the existing laws would permit in the protection and privileges of British subjects. (Library of the Society of Friends 1846b, 5th Month 16, 439).

Hodgkin's attention among Quakers in Britain immediately refocused on the need to promote and actively support education among aborigines:

The promotion of an improved education of their young people would be doubtless one of the most effectual means of opening their eyes to these advantages and of enabling them to reap the benefits within their reach. (Library of the Society of Friends 1846b, 5th Month 16, 439).

On 6 August 1847, the Committee on the Aborigines merged with the Africa Committee.

Thomas Hodgkin became a member of the merged groups, but it is here that we can leave the work of London Yearly Meeting and its Meeting for Sufferings and their philanthropic care for the plight of aborigines.

5 1834–1835: The Epistle from North Carolina

In this section and in the section ‘1838: Upper Canada’, we find traces of the Quaker Committee on the Aborigines directly encountering the relationships between colonists and aborigines (and ‘free people of colour’). This is at the heart of the work of this committee⁶⁰⁸ and that of the Parliamentary Select Committee on the Aborigines, as well as of Thomas Hodgkin’s life work as a champion for the rights of aborigines. It is at this time, in the 1830s, that Britain becomes more aware of, and is often shocked by, accounts reaching London about the treatment of non-colonists by colonists throughout the empire. This is the great matter⁶⁰⁹ that has brought together the people and their networks that are the subject of

⁶⁰⁸ This appendix will provide indications of concern for the relationships between colonists and others, for example see the advice to Quaker colonists in (Library of the Society of Friends 1839, 11th Month 1, 612): ‘We would entreat those who may establish themselves in newly settled countries to reflect upon the responsibility which attaches to them when they are the neighbours of uncivilised and heathen tribes. It is an awful but indisputable fact that most settlements of this description, besides dispossessing the natives of their lands without equivalent, have hitherto been productive of incalculable injury to the moral and physical condition of the native races which have been thereby more or less reduced in numbers, and in some instances completely exterminated.’

⁶⁰⁹ For a full analysis of the wave of change sweeping Britain in the wake of the 1832 reform of parliament, see (Heartfield 2011, 6 - 8).

this study, and the study will argue that it is this urgent need to reform how Britain and British colonists interact and deal with the colonial world that provides the forge in which British anthropology will be fashioned in the 1840s.

The next year, on 27 May 1834, London Yearly Meeting considered an Epistle⁶¹⁰ from North Carolina Meeting urging that North Carolina Meeting be allowed to evict 'coloured peoples' in that territory in return for 'some pecuniary assistance'.⁶¹¹ It is possible that this assistance was to be a compensation payment to coloured peoples for their displacement from North Carolina and into the Free States, and North Carolina Quakers were here asking for help in raising funds.

It may be that London Yearly Meeting and its Meeting for Sufferings were now becoming more aware of, and concerned about, the plight of aborigines (indicated in the previous year's report of the 'committee appointed to consider communicating to the heathen the knowledge of the Gospel', which ended up instead lamenting the plight of aborigines), and that this Minute from North Carolina required more consideration before making a response for the Meeting to consider; we see that a collection among Quakers in Britain was not immediately approved. Instead, the matter was taken away to be considered by the

⁶¹⁰ For an explanation of an 'Epistle' within the context of Quaker constitution, see the Fifth Edition of *Quaker faith & practice*, Chapter 6, 6.23 (<http://qfp.quaker.org.uk/passage/6-23>). (Accessed 23/5/2023)

⁶¹¹ 'An epistle has now been received and read from the Meeting for Sufferings for North Carolina held 14th of the 4th month, 1834 on the subject of the coloured population under the care of our Friends there, it is concluded to propose to the yearly meeting that liberty should be allowed to this meeting to render some pecuniary assistance on the occasion, with an especial view to the removal, with their own consent, of the said people of the State of North Carolina' (Library of the Society of Friends 1834, 5th Mon th 30, 448).

Committee on Epistles.⁶¹² Were Quakers in Britain possibly concerned about the needs of North Carolina Quaker colonists and a desire to treat ‘people of colour’ in North Carolina fairly? The following year (1835), the North Carolina matter of the displacement of ‘people of colour’ seems to have been resolved at London Yearly Meeting when Samuel Gurney reported back on behalf of the Committee on Epistles: ‘This meeting approves of the proceedings of the Meeting for Sufferings [to go ahead with a collection among all Quaker Meetings in Britain over the next two years] in regard to the coloured people under the care of the yearly meeting of North Carolina. Samuel Gurney, clerk’ (Library of the Society of Friends 1835, 5th Month 26, 523).

It should be noted that the British House of Commons Select Committee on the Aborigines was established in 1834 with the Quaker Joseph Pease among its membership. Interest in the relationships between colonists and aborigines was gathering momentum in the UK and the Epistle from North Carolina should also be viewed in this context.

6 1837: The Select Committee Report, the establishment of the Committee on the Aborigines and the Aborigines Protection Society

⁶¹² ‘The following minute has been brought in and read and this meeting recommend the subject to which it refers to the close attention of the Meeting for Sufferings with directions to make report thereon to the yearly meeting next year. Committee on Epistles 29th May 1834, This committee has been introduced into much concern on behalf of the aborigines of North America and suggests to the Yearly Meeting to recommend the subject to the close attention of the Meeting for Sufferings with directions to make report thereon. Samuel Gurney Clerk’ (Library of the Society of Friends 1834, 5th Month 30, 479).

In 1836, London Yearly Meeting did not minute matters relating to aborigines, but in 1837, the House of Commons Select Committee on the Aborigines published its report, largely condemning the widespread abuse of aborigines throughout the colonies (Aborigines Protection 1837). On 5 June 1837, the Meeting for Sufferings appointed a Committee on the Aborigines 'which is to be a subcommittee of the General Committee' (of the Meeting for Sufferings). It was a much smaller committee than its predecessor, the Committee of Enquiry, with at its inception only ten members (and four of these were previously members of the Committee of Enquiry).⁶¹³ The Committee on the Aborigines then quickly made a report to London Yearly Meeting through the General Committee that it had begun its work.⁶¹⁴

On 8 June 1837, almost immediately after the House of Commons Aborigines Select Committee report was published and three days after the establishment of the Quaker Committee on the Aborigines, Thomas Hodgkin, together with Sir Thomas Fowell Buxton MP (the Select Committee chair) and Joseph Pease, formed the Aborigines Protection Society at

⁶¹³ 'The following friends are appointed to take such steps as they may be enabled, to further the objects of the Yearly Meeting, as conveyed in its minute on the subject of the present state of the aborigines of the British Colonial Possessions, particularly as it respects those of the Indians of Upper Canada, viz William Allen, Thomas Christy, John Hodgkin Jun, John Thomas Barry, William Forster, Robert Forster, Abram Rawlinson Barclay, George Stacey, Robert Howard and Henry Knight Jun' (Library of the Society of Friends 1837b, 6th Month 5, 394).

⁶¹⁴ 'The following minute from the General Committee has been brought in and read. This meeting adopts the suggestion therein contained and desires the Meeting for Sufferings to pay close attention to the subject and to act in it at their discretion. General Committee 6th Mo 1st 1837. This committee having under its serious consideration the circumstances of the aborigines of the British Colonial possessions, particularly the Indians of upper Canada submits to the yearly meeting the propriety of recommending the close attention of the subject to the Meeting for Sufferings. Young Sturge, Clerk' (Library of the Society of Friends 1837a, 5th Month 31, 108).

Ratcliffe Quaker Meeting House in London. Thomas Hodgkin acted as secretary from at least 1844 (British and Foreign Aborigines' Protection 1844), and his brother John Hodgkin Jnr sat on the Quaker Committee on the Aborigines from 1837.

In July 1837, the Quaker Committee on the Aborigines was further strengthened with the addition of eighteen more members.⁶¹⁵ Meeting for Sufferings minute books noted that the Committee on the Aborigines met monthly.⁶¹⁶

The relationship between the Aborigines Protection Society and the Quaker Committee on the Aborigines is unclear simply from a reading of the Quaker corporate records. A survey of the publications of the two committees over the decade from 1837 to 1846 (see below) may help to reveal a relationship, as also will an analysis of membership lists of both groups (see Chapter 6), for example if a significant number of the members of the Quaker Committee on the Aborigines were also members of the Aborigines Protection Society.

There are gaps in the archived records of the Aborigines Protection Society (1838 and 1841–1846)⁶¹⁷ and an exhaustive search has not produced an indication that printed records of

⁶¹⁵ 'The friends appointed on the subject of the present state of the aborigines of the British Colonial Possessions are continued: the following friends are now added. Josiah Forster, William Hargreave, Samuel Darton, John Sanderson, Peter Bedford, Thomas Norton Jun, Richard Barrett, John Hamilton, Edwd Harris, Robert Alsop Jun, John Kitching, Joseph Storrs, Geo Holmes, Joseph Shewell, John Bell, Joseph Neatby, John Barclay, and Joseph Talwin Foster' (Library of the Society of Friends 1837b, 7th Month 7, 419-420).

⁶¹⁶ 'Josiah Forster reports that the committee on the state of the aborigines in the British Colonies met and attended to their appointments: the committee is continued' (Library of the Society of Friends 1837b, 8th Month 4, 429). 'The committee on the state of the aborigines in the British Colonies is continued' (Library of the Society of Friends 1837b, 9th Month 7, 539).

⁶¹⁷ See RAI Archive finding aid catalogue ABORIGINES' PROTECTION SOCIETY (A111).

business were indeed made by the Aborigines Protection Society in this period. Given the probable highly common membership of both organisations, this lack of Aborigines Protection Society record keeping might have been of no concern at the time if, as seems possible, the Aborigines Protection Society saw itself largely as a fellow traveller and cooperative producer of publications alongside the Quaker Committee on the Aborigines.

Over the next two years, with the publication of several reports in the public domain and a memorial to the Earl of Durham, the Governor General in America and Lord John Russell, the Secretary of State, the cause of aborigines was taken up once again by the Quaker Committee on the Aborigines.

7 1838: Upper Canada

James Heartfield has discussed the crisis in Upper Canada in detail (2011, 207 - 213). What follows is a precis of his account. Upper and Lower Canada had been in a state of tension since 1776, when refugees from the United States migrated north into Upper Canada during and then after the American War of Independence. Upper Canada was politically divided between nascent republicans and loyalists (loyal to the British Crown). Upper Canada was protestant and English speaking and Lower Canada Catholic and French speaking.⁶¹⁸

⁶¹⁸ The text is here referring to French-speaking Quebec.

Earlier in the decade, tensions had led to violent clashes between all factions over land rights and land use. The aborigines had been used as paid militia by all sides, but mostly on the side of the Crown, where their loyalties were felt to lie. In 1837, the Governor General John Lambton, 1st Earl of Durham, at the request of the British Parliament, made a tour of the region to report on the situation and to recommend a remedy. His report, delivered in 1838, recommended a form of self-rule for the region ((The Canadian Crisis and Lord Durham's Mission 1838).

At the same time the Governor of Upper Canada, Francis Bond Head, had invented his own solution to the land crisis by 'buying' land for settlers and his supporters from the aborigines and relocating them to Manitoulin Island, a considerable distance away. More than 500,000 people were displaced and the relocation later proved to be a disaster.

In May 1838, the Quaker Committee on the Aborigines produced a Memorial to the Earl of Durham, as well as a tract in which the crisis investigated by Lord Durham in Upper Canada is discussed at length. A few months later the Aborigines Protection Society also published a tract on the matter (Aborigines Protection Society 1839, 17 - 19). The concern may have been in part similar to that faced in 1834 by Quakers in Britain regarding North Carolina (the eviction of non-colonists out of settled territory).

During the next year, and up to June 1839, the Quaker Committee on the Aborigines seems to have been busy and relatively trouble free.⁶¹⁹

8 1839: Thomas Hodgkin – the Committee on the Aborigines and the memorial to Lord John Russell

In June 1839, Thomas Hodgkin joined the Quaker Committee on the Aborigines, appointed by the clerk of the Meeting for Sufferings, but we do not know why he had done so.⁶²⁰ There was, however, a noticeable quickening of the pace of activity within the committee from here onwards. The first act of the committee that included Thomas Hodgkin was to address the behaviour of Quakers in the colonies towards aborigines in a minute offered to the Meeting for Sufferings. We can speculate that Hodgkin had been brought in to help direct

⁶¹⁹ 'Some information on the state of the aborigines of the British Colonial possessions has been laid before this meeting by the Meeting for Sufferings which is desired to bring to the next yearly meeting such further information as it may receive on this subject. This meeting has been deeply interested with some information now communicated respecting the circumstances of the African race, both in their native land and in the colonies of this country; and understanding that much pecuniary assistance will be required to promote their welfare as respects their education in our colonies and otherwise this meeting thinks it right to encourage a very liberal subscription among our members throughout the country to be applied under the direction and at the discretion of the Meeting for Sufferings in connection with the object above mentioned' (Library of the Society of Friends 1838a, 5th Month 29, 155-156). The Committee on the Aborigines had, it seems, been very active, as is evidenced by the level of intelligence and flows of money from collections that it was now regularly feeding into London Yearly Meeting and the Meeting for Sufferings. Robert Forster reported that the committee appointed on the subject of the Aborigines had met since the previous month and the committee continued (Library of the Society of Friends 1838b, 10th Month 5, 542). On 29 May 1839, the minutes of London Yearly Meeting recorded a continuing interest in the welfare of aborigines but did not propose any further measures then, other than that the Meeting for Sufferings was 'encouraged to continue its attention to the subject and to transmit to our next yearly meeting any further information which it may obtain'.

⁶²⁰ 'The list of correspondents has now been called over, and the following alterations and additions have been proposed and agreed to viz; "In London , Thomas Hodgkin added"' (Library of the Society of Friends 1839, 6th Month 7, 587-588).

the Committee in the light of what the Committee had now learned from the Earl of Durham's report on Upper Canada, which had exposed the behaviour of colonists, and especially their representative the Colonial Governor, Francis Bond Head. The Committee was particularly exercised over the reports of mistreatment of aborigines under forced migration.⁶²¹ This, taken with the Committee's prior experience in 1834 with North Carolina, may have incentivised it to act to put its own house in order with respect to the conduct of Quakers. Thomas Hodgkin may have simply been seen as the best man for the job.⁶²² More research is needed to properly analyse the events of these years.

⁶²¹ The Committee on the Aborigines published a twenty-four-page tract consisting of six separate reports of mistreatment (Library of the Society of Friends. LYM Meeting for Sufferings 1839a).

⁶²² 'A minute has been brought in from the committee on the Aborigines for the adoption of this meeting which has been considered, and is, with a small alteration agreed to, the minute is as follows, copies of the same are to be forwarded to the different Quarterly Meetings. The case of members of our society who may be contemplating emigration to distant colonies has been brought under the notice of this meeting; and we think it right to remind our dear friends who may be so circumstanced how much the steps they take make affect not merely the interests of humanity but moral and Christian principle. We would entreat those who may establish themselves in newly settled countries to reflect upon the responsibility which attaches to them when they are the neighbours of uncivilised and heathen tribes. It is an awful but indisputable fact that most settlements of this description, besides dispossessing the natives of their lands without equivalent, have hitherto been productive of incalculable injury to the moral and physical condition of the native races which have been thereby more or less reduced in numbers, and in some instances completely exterminated. Earnestly therefore do we desire that all those under our name who may emigrate to such settlements may be careful neither directly nor indirectly to inflict injury upon the natives, but that they may on the contrary in their whole conduct exhibit the practical character of that religion that breathes peace on earth and goodwill toward men. As this is their aim they will not only exert themselves to check the evils which are but too generally inflicted by the whites upon their feeblers neighbours but will be solicitous to do their part in endeavouring to diffuse amongst them the blessings of civilisation and Christianity, which will prove the best means of preventing their extermination, and of raising them to the full enjoyment of their rights' (Library of the Society of Friends 1839, 11th Month 1, 612).

Next, the committee's attention moved to Van Diemen's Land and the committee set about drafting a Memorial to Lord John Russell, Secretary of State for the Colonies.⁶²³ We can here perhaps see a twin-track approach if we take the Memorial to be a complementary action alongside that of providing advice to Quaker colonists.⁶²⁴ We can also note here that, once again, the actions taken by the Quaker Committee on the Aborigines was reinforced by parallel action taken contemporaneously by the Aborigines Protection Society (Heartfield 2011, 91). In February of the following year, 1840, the Meeting for Sufferings recorded that 'Samuel Gurney reports that the Memorial to Lord John Russell on the subject of the aborigines was presented to him on the 8th of last month' (Library of the Society of Friends 1840b, 2nd Month 7, 5). The Committee also published a tract (*An Address of Christian counsel and caution to emigrants to newly-settled colonies* 1841).

During the remainder of 1840 and 1841, the Committee on the Aborigines continued to meet regularly and to circulate documents. In August, the committee reorganised once

⁶²³ The Memorial, which is a general plea for the universal sufferings of aborigines, made this request: 'We allude to the recognition and security of their title to some portion of the territories once wholly theirs, to the Bona fide admission of their evidence in courts of law, to the recognition of their rights as men and citizens, to a full participation in all of the privileges of British subjects, so that the distinctions of colour and race may no longer operate against them and that effectual steps may be taken both at home and in the colonies to effect their elevation, in a moral, intellectual and political point of view' (Library of the Society of Friends. LYM Meeting for Sufferings 1840).

⁶²⁴ 'A Memorial has been produced and read from the committee on the Aborigines addressed to Lord John Russell as secretary of state for the colonies, which the said committee suggests for the adoption of this meeting, the memorial has been several times read and considered, and is with a few alterations been adopted and signed by the clerk on behalf of the meeting, the case of presenting the same is left to the following friends, viz. William Allen, Samuel Gurney, Thomas Hodgkin, Robert Forster, George Stacey and John Hodgkin Jun' (Library of the Society of Friends 1839, 12th Month 6, 616).

again, but the reason for this is not clear.⁶²⁵ It does however present us with another list of names of committee members. Chapter 6 analyses these names of members within the scope of a small network analysis.

9 1842: The enlargement of the Committee on the Aborigines

In 1842, doubt began to emerge about the continuance of the Committee on the Aborigines.⁶²⁶ There is only one small hint in the Meeting for Sufferings records about the committee questioning its further continuance, and that is when the committee asked that,

⁶²⁵ 'This list of the aborigines committee has been revised and it is agreed that it now consist of the following friends, William Allen, John Hodgkin Jun, John Thomas Barry, William Forster, Robert Forster, George Stacey, Henry Knight, Josiah Forster, John Sanderson, Peter Bedford, Thomas Norton Jun, Edward Harris, Robert Alsop Jun, Joseph Storrs, Joseph Talvin Foster, William Grimshaw Jun, Edward Paull and Thomas Hodgkin' (Library of the Society of Friends 1840b, 8th month 7, 48).

⁶²⁶ 'The following report was received from the committee on the aborigines. At a committee on the aborigines held on 7th of the 10th month 1842. The committee of the Meeting for Sufferings on the subject of the aborigines in considering (before the occurrence of the last yearly meeting) the propriety of presenting a report seriously entertained the question whether it might not be right to propose the discontinuance of the committee altogether, but remembering the interest which had been from time to time expressed in the yearly meeting on the subject, and observing in a retrospect of the committee's labours in previous years that several subjects of practical importance had come before them and also before the Meeting for sufferings in connection with the wrongs of the aborigines, they are reluctant to take this course. They have met since the yearly meeting and have again looked both at the general question and also at some of the branches of it which they apprehend that a body acting on behalf of the society could most advantageously entertain. At the same time they are desirous that the Meeting for Sufferings should devote a little time to the serious consideration of the whole subject; and decide for itself what is most advisable to be done in the interval between the present time and the occurrence of the next yearly meeting; and that if any committee be continued the meeting would revise the present list. They also submit to the Meeting for Sufferings that if a committee be continued that all papers received by the meeting bearing on the interests of the aborigines should be referred to such committee for attention, and that to it should be the examination and printing of the materials of this description which were presented to the last yearly meeting. A copy of this minute to be taken to the Meeting for Sufferings' (Library of the Society of Friends 1840b, 8th Month 7, 56).

if it were to continue, then the Committee on the Aborigines alone should be in receipt of all correspondence about aborigines, and also be solely responsible for publications. The Committee on the Aborigines may have felt itself in conflict with the Africa Committee.⁶²⁷ It is also possible that Thomas Hodgkin was becoming dissatisfied with the work of the Quaker Committee when compared to that of the Aborigines Protection Society, and perhaps he began to see that as a successor organisation. Again, more research is required.

The Meeting for Sufferings responded by requesting that the Committee continue in its work, and further increased its membership by an additional twenty-two new members. Two years later (in 1844) the Meeting for Sufferings called for the Committee to report to the upcoming Yearly Meeting.⁶²⁸

10 1845–1846: The end of the Committee on the Aborigines

In 1845 the work of the Committee on the Aborigines continued, now with Thomas Hodgkin speaking for the committee. We can sense that in spite of diligently holding regular

⁶²⁷ On 16 August 1847 the Quaker Committee on the Aborigines merged with the Africa Committee.

⁶²⁸ 'This meeting agrees to the suggestion in the report and desires to encourage the committee to continue their attention to the subject under their care in relation to their appointment. James Bowden, William Nash, and Joseph Sturge are now added to the committee' (Library of the Society of Friends 1842, 10th Month 7, 195-197). 'The committee on the state of the aborigines are requested to present a brief report of their proceedings during the past year to the approaching yearly meeting' (Library of the Society of Friends 1844, 5th Month 3, 286).

meetings, the Committee now felt overwhelmed with the claims made upon it, for example in North America, Guyana, Australia and New Zealand.⁶²⁹

The Committee continued to have small successes, however, especially in its influence (probably via the colonial Aborigines Protection Society) in New Zealand.⁶³⁰ Also, 'The committee has published a small pamphlet consisting of the two reports of the preceding years, with an appendix containing information respecting the Indians of British North America, as well as those tribes more immediately under the notice of the American yearly meetings' (Library of the Society of Friends 1845, 5th Month 16, 345-346).

In addition to the inevitable strains of coping with the task of influencing colonists and the Colonial Office on behalf of aborigines, the Committee increasingly had to contend with

⁶²⁹ 'The following report has been presented by the Aborigines Committee. The Aborigines Committee of the Meeting for Sufferings has with scarcely any interruption met monthly during the past year in the course of which it has received information respecting the aborigines in several of the British Colonies and also from parts which are independent of the country. With little exception these accounts have exhibited the feeble, the ignorant and the pagan suffering in various ways from the oppressive hands of their civilised fellow creatures who abusing the superiority conferred by the knowledge and resources which it is their privilege to have received, dishonour the Christian name which they unworthily profess. Amongst these cases may be mentioned numerous encroachments on the lands and other possessions of North American Indians – the kidnapping of Indians of Guyana by Brazilian slave dealers – the deportation of negroes and coolies under the designation of free emigrants – the destruction of Australian natives by means of poison introduced into their food, which is perpetrated with impunity not from the inhumanity or indifference of the officers of government but from the defect of the law which gives no validity to the testimony of aborigines, and the efforts which are being made to obtain possession of land in New Zealand in violation of the rights of the native population' (Library of the Society of Friends 1845, 5th month 16, 328).

⁶³⁰ 'Although these circumstances are far from evincing any improvement in the state of things in relation to the subject, the committee has not seen its way to take any particular steps, except that very recently individual members of the committee have used some exertion in behalf of the right of the natives of New Zealand' (Library of the Society of Friends 1845, 5th Month 16, 342).

public apathy.⁶³¹ There was also a growing awareness that the publications of the Committee were no longer popular.⁶³² In this context Thomas Hodgkin made one last plea to the Meeting for Sufferings:

In conclusion the committee would again recommend the subject for the increased attention and sympathy of Friends.

Signed on behalf of the committee.

Devonshire House 14th of the 5th Month 1845.

Thomas Hodgkin. (Library of the Society of Friends 1845, 5th Month 16, 345-346).

There may even have been resource allocation issues between the Africa Committee and the Committee on the Aborigines because funds were divided up between them, as noted at Yearly Meeting in 1846.⁶³³

On 16 May 1846, the Committee on the Aborigines made its last report to the Meeting for Sufferings, lamenting that 'Nothing has occurred since the last yearly meeting to warrant

⁶³¹ 'During the past year the committee has been made sensible that whilst there exists numerous cases demanding sympathy, coming almost within the range of personal observation, it is extremely difficult to sustain, or even to excite an interest in behalf of those distant branches of the human family for whose sake this committee has been appointed' (Library of the Society of Friends 1846a, 5th Month 26, 582).

⁶³² 'The committee was apprehensive that the pamphlets relating to the aborigines already published by direction of the Meeting for sufferings are far from being generally diffused amongst the members of the Society, and that even the libraries of meetings are not completely supplied with them' (Library of the Society of Friends 1846a, 5th Month 26, 599).

⁶³³ 'A report has now been brought in from the Meeting for Sufferings as to the amount received from the different Quarterly Meetings in conformity with the minute of last year recommending a subscription on behalf of the coloured population in the West India Islands and elsewhere and the aborigines of different countries. The said report contains also information as to the appropriation of a part of the said fund. It is satisfactory to this meeting which desires the Meeting for Sufferings to transmit annual to this meeting the state of the fund' (Library of the Society of Friends 1846a, 5th Month 26, 606).

the belief that any important changes have been made to advance the moral and physical improvement of the aborigines in the British colonies, or that they are enjoying to the full extent what the existing laws would permit in the protection and privileges of British subjects' (Library of the Society of Friends 1846b, 5th Month 16, 439).

On 6 August 1847, the Committee on the Aborigines merged with the Africa Committee and Thomas Hodgkin became a member of the merged groups. From now on Hodgkin, and the Quakers in Britain who shared his passion for the plight of aborigines, continued their labours within the Aborigines Protection Society. Within London Yearly Meeting's committee work and the Africa Committee, Hodgkin refocused on the need to promote and actively support education among aborigines.⁶³⁴

11 Lists of names of Quaker committee members

Several records of the London Yearly Meeting and its Meeting for Sufferings throughout the period examined reveal lists of names of members of the various committees considered here:

1832 – The Committee of Enquiry, 55 members.

⁶³⁴ 'The promotion of an improved education of their young people would be doubtless one of the most effectual means of opening their eyes to these advantages and of enabling them to reap the benefits within their reach' (Library of the Society of Friends 1846b, 5th month 16, 439).

1837 – The Committee of the Aborigines, 10 members.

1837 – Later that year, members increased to 28.

1839 – Thomas Hodgkin added, now 29 members.

1840 – Numbers reduced, 18 members.

1842 – Numbers increased, 21 members.

From an examination of the Quaker corporate manuscripts here, we can derive possible lines of further enquiry that small network analysis may help to resolve:

- To what extent did the Quaker practice of endogamy or ‘marrying in’ support the building and maintaining of strong social networks?
- Many of the named supporters of the organisation that led to the institution of anthropology in Britain were Doctors of Medicine. How do the social networks of occupation compare to the social networks of religion (here in the form of Quakers)?

12 Conclusion

An examination of the corporate records of Quakers in Britain has found evidence of the Quaker Committee on the Aborigines (and its precursor the Committee of Enquiry) and publications produced by the Meeting for Sufferings on behalf of the Committee in spite of

no records of the proceedings of the Committees themselves surviving. This evidence has revealed:

- The origins of the Quaker Committee on the Aborigines and its remit, and that its remit was largely in sympathy with that of the Aborigines Protection Society.
- The roles played in or alongside the Quaker Committee on the Aborigines by members of the Aborigines Protection Society, especially Thomas Hodgkin, Joseph and Samuel Gurney and Joseph Pease.

It has been discovered that:

- The Committee of Enquiry that immediately preceded the Quaker Committee on the Aborigines, while seeking initially to address an evangelical concern, found that concern quickly transformed into a philanthropic one and this in turn may have resulted from opening up Quaker networks linking London to the colonies; those networks may have provided a conduit along which information and concerns about the plight of aborigines flowed.⁶³⁵
- The Quaker Committee on the Aborigines formed as a response to the emerging concern for the plight of aborigines included John Hodgkin and later his brother Thomas Hodgkin, who three days after the formation of the Quaker Committee on

⁶³⁵ See (Laidlaw 2005b) for a detailed analysis of how social networks worked at this time, especially between London and the colonies.

the Aborigines then formed the secular Aborigines Protection Society at Ratcliffe Quaker Meeting House. This suggests a high degree of coordination between the brothers.

- In the period 1837–1846, both the Quaker Committee on the Aborigines and the secular Aborigines Protection Society lobbied and published on the same matters, often seemingly in cooperation with each other.
- The Quaker corporate records analysed above include several list of names that can be taken up into a small network analysis in Chapter 6.

In summary, the manuscript review of Quaker corporate records reveals a Quaker Committee on the Aborigines that indeed fits within the group of nascent organisations that this research takes to be those leading to the formation of the Ethnological Society of London and the institution of anthropology in Britain.

13 Publications of the Quaker Committee on the Aborigines

The Quaker Committee on the Aborigines produced nine pamphlets between the years 1838 and 1845 ranging over issues such as:

- Information respecting the Aborigines in British Colonies 1838 (Library of the Society of Friends. LYM Meeting for Sufferings 1838)
- Extracts from the Proceedings from the Committee on Indian Affairs 1839 (Library of the Society of Friends. LYM Meeting for Sufferings 1839b)
- Committee on Indian Affairs 1843 (Library of the Society of Friends. LYM Meeting for Sufferings 1843)
- Facts relating to the Canadian Indians 1839 (Library of the Society of Friends. LYM Meeting for Sufferings 1839a)
- Address to Lord John Russell 1840 (Library of the Society of Friends. LYM Meeting for Sufferings 1840)
- An address of Christian counsel and caution to emigrants 1841
- Report on Meeting for Sufferings on Aborigines 1841 (Library of the Society of Friends. LYM Meeting for Sufferings 1841)

- Reports of the Committee on Indian Affairs Philadelphia 1842 (Library of the Society of Friends. LYM Meeting for Sufferings 1842)
- Civilisation and Instruction of the Indians 1844 (Library of the Society of Friends. LYM Meeting for Sufferings 1844)

Appendix 4 Records in Context. Definition of terms – Person

Agent to Agent relations [e.g., Persons]. Any relation that holds between an agent and another agent. (International Council on Archives 2023, 76)

RiC-E08	Person
Attribute ID	Attribute Name
RiC-A43	General Description
RiC-A22	Identifier
RiC-A28	Name
RiC-A21	History
RiC-A25	Language
RiC-A26	Legal Status
RiC-A15	Demographic Group
RiC-A30	Occupation Type

RiC E08 Definition of Terms -Person (International Council on Archives 2023, 70)

ID	RiC-E08
Name	Person
Definition	An individual human being.

Scope Notes	<p><i>Person</i> is a kind of Agent (RiC-E07).</p> <p>Most commonly, a human being (biological person) has a single socially constructed identity or persona.</p> <p>Less common though not rare, one or more personae in addition to the original persona which emerges at or near birth may be associated with the human being over the course of that human being's lifetime. Such "alternative personae" are most often created by the original <i>person</i> for specific purposes. The original persona may, in everyday discourse, be regarded as "the real person."</p> <p>Under some circumstances, an alternative persona might eclipse or replace the original <i>person</i> (Mark Twain eclipsing Samuel Clemens; John Wayne eclipsing Marion Mitchell Morrison), that is, the alternative identity becomes the predominant identity.</p> <p>Less common is when two or more <i>persons</i> collaborate to create a shared persona. A persona shared by two or more persons constitutes a kind of <i>group</i>.</p> <p>Within the archival context, the description of a <i>person</i> commonly will focus on the original associated persona, with alternative personae noted. Exceptionally, an alternative persona may displace the original persona as the focus of the description.</p> <p>Under some circumstances, for example, when <i>record resources</i> are associated with two or more different personae of one <i>person</i>, describing the different personae as separate though related <i>persons</i> may be desirable.</p> <p>Alternatively, a <i>person</i> may change their identity over the course of their lifetime.</p>
Examples	<p>Nelson Mandela [activist, politician]</p> <p>Jean Harlow [actress]</p>
Comments	

RiC E08 – Definition of terms – Person (International Council on Archives 2023, 28)

ID	RiC-R017	
Name	<i>has descendant</i>	inverse relation: <i>has ancestor</i>
Domain/Range	Person	Person
Cardinality	M to M	
Definition	Connects a <i>person</i> to one of their descendants.	
Scope Notes	There may be zero to many intermediate <i>persons</i> , ignored or unknown, between the two connected <i>persons</i> .	
Examples	<p>Marc Ferrez <i>has descendant</i> Gilberto Ferrez. [pt]</p> <p>Gilberto Ferrez <i>has ancestor</i> Marc Ferrez. [pt]</p>	
Relation types	<p>Sequential relations</p> <p>Agent to agent relations</p>	
Broader relations	<p>RiC-R016 <i>has successor</i></p> <p>RiC-R047 <i>has family association with</i></p>	
Narrower relations	RiC-R018 <i>has child</i>	

RiC R017 – Definition of terms – Person (E08) has descendant / ancestor (International Council on Archives 2023, 90)

ID	RiC-R018	
Name	<i>has child</i>	inverse relation: <i>is child of</i>
Domain/Range	Person	Person
Cardinality	M to M	
Definition	Connects a <i>person</i> to one of their children.	
Scope Notes		
Examples	<p>Alfonso Carlos de Borbón y Austria-Este (1849-1936) <i>is child of</i> M^a Beatriz de Austria-Este (1824- 1906). [es]</p> <p>Júlio Ferrez <i>has child</i> Gilberto Ferrez. [pt]</p> <p>Gilberto Ferrez <i>is child of</i> Júlio Ferrez. [pt]</p>	
Relation types	<p>Sequential relations</p> <p>Agent to agent relations</p>	
Broader relations	RiC-R017 <i>has descendant</i>	
Narrower relations	None	

RiC R018 – Definition of terms – Person (E08) has child / is the child of. (International Council on Archives 2023, 91)

Appendix 5 Archival Records Standards – ISAAR – CPF Second Edition 2004

<https://www.ica.org/reSource/isaar-cpf-international-standard-archival-authority-record-for-corporate-bodies-persons-and-families-2nd-edition/> (Accessed 16 June 2024)

1.6 Where a number of repositories hold records from a given source they can more easily share or link contextual information about this source if it has been maintained in a standardized manner. Such standardization is of particular international benefit when the sharing or linking of contextual information is likely to cross national boundaries. The multinational character of past and present record keeping creates the incentive for international standardization which will support the exchange of contextual information. For example, processes such as colonialization, immigration and trade have contributed to the multinational character of recordkeeping.

Archival authority records are similar to library authority records in as much as both forms of authority record need to support the creation of standardized access points in descriptions. The name of the creator of the unit of description is one of the most important of such access points. Access points may rely on the use of qualifiers that are deemed essential to clarify the identity of the entity thus named, so that accurate distinctions may be made between different entities that have the same or very similar names.

1.8

The primary purpose, therefore, of this standard is to provide general rules for the standardization of archival descriptions of records creators and the context of records creation, thus enabling:

- access to archives and records based on the provision of descriptions of the context of records creation that are linked to descriptions of the often diverse and physically dispersed records themselves;
- understanding by users of the context underlying the creation and use of archives and records so that they can better interpret their meaning and significance;
- precise identification of records creators incorporating descriptions of relationships between different entities, especially documentation of administrative change within corporate bodies or personal change of circumstances in individuals and families; and

- the exchange of these descriptions between institutions, systems and/or networks. 1.10

Authority record. The authorized form of name combined with other information elements that identify and describe the named entity and may also point to other related authority records.

Record. Information in any form or medium, created or received and maintained by an organization or person in the transaction of business or the conduct of affairs.

Rules and conventions for standardizing access points may be developed nationally or separately for each language. Vocabularies and conventions to be used in creating or selecting the data content for these elements may also be developed nationally, or separately for each language.

This standard addresses only part of the conditions needed to support the exchange of archival authority information. Successful automated exchange of archival authority information over computer networks is dependent upon the adoption of a suitable communication format by the repositories involved in the exchange. Encoded Archival Context (EAC) is one such communications format which supports the exchange of ISAAR(CPF) compliant archival authority data over the World Wide Web. EAC has been developed in the form of Document Type Definitions (DTDs) in XML (Extensible Markup Language) and SGML (Standard Generalized Markup Language).

Authorized form(s) of name

Purpose:

To create an authorized access point that uniquely identifies a corporate body, person or family.

Rule:

Record the standardized form of name for the entity being described in accordance with any relevant national or international conventions or rules applied by the agency that created the authority record. Use dates, place, jurisdiction, occupation, epithet and other qualifiers as appropriate to distinguish the authorized form of name from those of other entities with similar names. Specify separately in the

Rules and/or conventions element (5.4.3) which set of rules has been applied for this element. 5.12

Dates of existence

Purpose:

To indicate the dates of existence of the corporate body, person or family.

Rule:

Record the dates of existence of the entity being described. For corporate bodies include the date of establishment/foundation/enabling legislation and dissolution. For persons include the dates or approximate dates of birth and death or, when these dates are not known, *floruit* dates. Where parallel systems of dating are used, equivalences may be recorded according to relevant conventions or rules.

Specify in the Rules and/or conventions element (5.4.3) the system(s) of dating used, e.g. ISO 8601. 5.2.1

Mandates/Sources of authority

Purpose:

To indicate the Sources of authority for the corporate body, person or family in terms of its powers, functions, responsibilities or sphere of activities, including territorial.

Rule:

Record any document, law, directive or charter which acts as a source of authority for the powers, functions and responsibilities of the entity being described, together with information on the jurisdiction(s) and covering dates when the mandate(s) applied or were changed. 5.2.6

Internal structures/Genealogy

Purpose:

To describe and/or represent the internal administrative structure(s) of a corporate body or the genealogy of a family.

Rules:

Describe the internal structure of a corporate body and the dates of any changes to that structure that are significant to the understanding of the way that corporate body conducted its affairs (e.g. by means of dated organization charts).

Describe the genealogy of a family (e.g. by means of a family tree) in a way that demonstrates the inter-relationships of its members with covering dates. 5.2.7

Names/Identifiers of related corporate bodies, persons or families

Purpose:

To indicate the names and any unique identifiers of related entities and to support linkages to the authority records for related corporate bodies, persons or families.

Rule:

Record the authorized form of name and any relevant unique identifiers, including the authority record identifier, for the related entity. 5.3.1

5.3.2 Category of relationship

Purpose:

To identify the general category of relationship between the entity being described and another corporate body, person or family.

Rule:

Record a general category into which the relationship being described falls. Use general categories prescribed by national rules and/or conventions or one of the following four categories. Record in the

Rules and/or conventions element (5.4.3) any classification scheme used as a source of controlled

vocabulary terms to describe the relationship.

- **Hierarchical** (e.g. superior/subordinate; controlled/controlling; owner of/owned by)

In a hierarchical relationship an entity may exercise some authority and control over the activities of a number of other corporate bodies, persons or families. An entity may also be subordinate to a number of other corporate bodies, persons or families, as for example a joint-committee or an organization whose superior changed over time.

- **Temporal** (e.g. predecessor/successor)

In a temporal relationship an entity may succeed a number of other corporate bodies, persons or families in exercising some functions and activities. In turn it may be succeeded by a number of other corporate bodies, persons or families.

- **Family**

In a family a person may have a wide circle of relationships with other members of the family and with the family as an entity. Where the genealogical structure of the family is complex it may be appropriate to create separate authority records for each member and link them to parent(s),

spouse(s) and child(ren). Alternatively this information may be recorded in the Internal structures/Genealogy element

Authority record identifier

Purpose:

To identify the authority record uniquely within the context in which it will be used.

Rule:

Record a unique authority record identifier in accordance with local and/or national conventions. 5.4.1

Sources

Purpose:

To identify the Sources consulted in creating the authority record.

Rule:

Record the Sources consulted in establishing the authority record.

Examples:

HMC, *Principal Family and Estate Collections: Family Names L-W*, 1999

Complete Peerage, 1936

Burkes Peerage, 1970

Complete Baronetage, vol 5, 1906

United Kingdom, *The National Archives: Historical Manuscripts Commission*. 5.4.8

Archival authority records are created primarily to document the context of records creation. To make this documentation useful it is necessary to link the authority records to descriptions of records.

Archival authority records can also be linked to other relevant information resources . When such linkages are made it is important to describe the nature of the relationship, where known, between the corporate body, person or family and the linked reSource. This Section provides guidance on how such linkages can be created in the context of an archival descriptive system. See Figure 1 for a pictorial

representation of this.

Example 5 - Person description

Language of description: English (Australia)

5.1 IDENTITY AREA		
5.1.1 Type of entity		Person
5.1.2 Authorized form of name		Mabo, Eddie, 1936-1992
5.1.5 Other forms of name		Mabo, Edward Koiki, 1936-1992
5.2 DESCRIPTION AREA		
5.2.1 Dates of existence		1936-1992
	<i>Dates ISO 8601</i>	1936/1992-01-21

5.2.2 History	<p>29 June 1936 - Born on Mer, the son of Robert Zezou Sambo and Annie Mabo of the Piadaram clan. Because his mother died in childbirth, he was adopted under customary law by his uncle Benny Mabo and aunt Maiga.</p> <p>1953-57 - Worked on trochus fishing luggers out of Mer.</p> <p>1957 - Left Mer and moved to the mainland. Worked at various jobs including canecutter and railway labourer.</p> <p>1959 - Married Bonita Nehow (born 1943).</p> <p>1960-61 - Union representative, Townsville-Mount Isa rail construction project.</p> <p>1962-67 - Worked for the Townsville Harbour Board. 1962-69 - Secretary, Aboriginal and Torres Strait Islander Advancement League.</p> <p>1967 - Helped organise seminar in Townsville: 'We the Australians: What is to Follow the Referendum?'</p> <p>1967-71 - Worked as gardener-groundsman, James Cook University</p> <p>1973 - Mabo and family travelled to Thursday Island en route to Mer with the intention of visiting Mabo's dying father, but were denied entry to Mer.</p> <p>1973-83 - Director, Black Community School, Townsville.</p> <p>1974-78 - Member of the Aboriginal Arts Council.</p> <p>1975-80 - President, Yumba Meta Housing Association.</p> <p>1975-78 - Member, National Aboriginal Education Committee. 1978-81 - Assistant Vocational Officer, Aboriginal Employment and Training Branch Commonwealth Employment Service.</p> <p>1978-79 - Member, Australian Institute of Aboriginal Studies Education Advisory Committee.</p> <p>1981-84 - Pursued Diploma of Teaching, Townsville College of Advanced Education/James Cook University.</p> <p>1981 - Conference on land rights at James Cook University. Decision to take the Murray Islanders' land case to the High Court 1982 - Land rights case launched. Plaintiffs were Mabo, Sam Passi, Father Dave Passi, James Rice and Celuia Mapo Salee.</p> <p>1986-87 - Director, ABIS Community Cooperative Society Ltd, Townsville.</p> <p>1986-87 - Assistant Director, Aboriginal Arts, Melbourne Moomba Festival.</p> <p>1987-88 - Employed by the Department of Aboriginal Affairs as Community Arts Liaison Officer, 5th Festival of Pacific Arts, Townsville.</p> <p>1987-88 - Vice-Chairman, Magani Malu Kes.</p> <p>1988 - High Court ruled the <i>Queensland Coast Islands Declaratory Act</i> 1985 contrary to the Commonwealth <i>Racial Discrimination Act</i> 1975.</p> <p>21 Jan. 1992 - Edward Koiki Mabo died in Brisbane.</p> <p>3 June 1992 - High Court delivered a 6:1 verdict in favour of Mabo, <i>Mabo v State of Queensland (No. 2)</i> (1992) 175 CLR 1, overturning the 205-year-old legal doctrine of <i>terra nullius</i>.</p> <p>26 Jan. 1993 - <i>The Australian</i> announced Eddie Mabo its 1992 Australian of the Year.</p>
5.2.3 Places	Mer [Murray Island], Torres Strait (1936-1957) Townsville, Queensland (c.1960-1992)
5.2.5 Functions, occupations and activities	<p>Trochus fisherman</p> <p>Sugarcane cutter Railway labourer Trade union official Waterfront worker</p> <p>Indigenous community leader Gardener</p> <p>Vocational officer</p> <p>Teacher</p> <p>Legal aid officer</p> <p>Indigenous arts administrator Indigenous land rights plaintiff</p>

5.2.6 Mandates Sources of authority		Torres Strait customary law
5.2.8 General context		Edward Koiki Mabo was born in 1936 on the island of Mer, one of the Murray Islands, which are located at the eastern extremity of Torres Strait. In June 1992, six months after his death, Mabo achieved national prominence as the successful principal plaintiff in the landmark High Court ruling on native land title. The High Court ruling, for the first time, gave legal recognition to the fact that indigenous land ownership existed in Australia before European settlement and that, in some cases, this land tenure was not subsequently extinguished by the Crown.
5.3 RELATIONSHIPS AREA		
<i>First Relation</i>		
5.3.1 Name / identifier of the related entity	<i>Authorized form of name</i>	Mabo, Bonita, 1943-
	<i>Other form of name</i>	Nehow, Bonita, 1943-
5.3.2 Category of relationship		Family
5.3.3 Description of relationship		Spouse
5.3.4 Dates of the relationship		1959-1992
	<i>Dates ISO 8601</i>	1959/1992-01-21
<i>Second Relation</i>		
5.3.1 Name / identifier of the related entity	<i>Authorized form of name</i>	Aboriginal and Torres Strait Islander Advancement League
5.3.2 Category of relationship		Associative
5.3.3 Description of relationship	<i>Title</i>	Secretary.
	<i>Narrative</i>	Mabo resigned from the League because of the involvement of people he considered to be insincere 'do-gooders'. He then established the all-black Council for the Rights of Indigenous People
5.3.4 Dates of the relationship		1962-1969
	<i>ISO 8601</i>	1962/1969
<i>Third Relation</i>		
5.3.1 Name / identifier of the related entity	<i>Authorized form of name</i>	Black Community School, Townsville, Qld
5.3.2 Category of relationship		Associative
5.3.3 Description of relationship	<i>Title</i>	Director
	<i>Narrative</i>	Mabo was Director of this School, the first of its kind established in Australia, throughout the ten years of its existence. The School, which was an independent school funded by the Commonwealth, was forced to close in 1983 because the lease on its site had expired and the School was unable to secure another site.
5.3.4 Dates of the relationship		1973-1983
	<i>ISO 8601</i>	1973/1983
<i>Fourth Relation</i>		
5.3.1 Name / identifier of the related entity	<i>Authorized form of name</i>	James Cook University of North Queensland
5.3.2 Category of relationship		Associative
		Employee

5.3.3 Description of relationship	<i>Title</i>	Gardener-Groundsman
5.3.4 Dates of the relationship		1967-1971
	<i>ISO 8601</i>	1967/1971
<i>Fifth Relation</i>		
5.3.1 Name / identifier of the related entity	<i>Authorized form of name</i>	James Cook University of North Queensland
	<i>Predecessor</i>	Townsville College of Advanced Education
5.3.2 Category of relationship		Associative
5.3.3 Description of relationship		Student
	<i>Narrative</i>	Mabo enrolled in a Diploma of Teaching course at Townsville College of Advance Education in 1981. In 1982, the College of Advanced Education amalgamated with the James Cook University of North Queensland. Mabo eventually decided not to become a teacher because he felt he was unsuited to classroom situations.
5.3.4 Dates of the relationship		1981-1984
<i>Sixth Relation</i>		
5.3.1 Name / identifier of the related entity	<i>Authorized form of name</i>	Council for the Rights of Indigenous People
5.3.2 Category of relationship		Associative

5.3.3 Description of relationship	<i>Title</i>	President
	<i>Narrative</i>	Established in 1970 as a break away from the Aboriginal and Torres Strait Islander Advancement League, this all-black Council established a legal aid service, a medical service and the Black Community School in Townsville.
5.3.4 Dates of the relationship		1970-c.1983
<i>Seventh Relation</i>		
5.3.1 Name / identifier of the related entity	<i>Authorized form of name</i>	Yumba Meta Housing Association
5.3.2 Category of relationship		Associative
5.3.3 Description of relationship	<i>Title</i>	President
	<i>Narrative</i>	The Yumba Meta Housing Association acquired houses in Townsville using Commonwealth funds and was responsible for renting them to black tenants. Mabo was President of the Association, 1975-80. During the period 1978-80, Mabo's presidency was contested by a group of disaffected members and evicted tenants who formed a new Board of Directors.
5.3.4 Dates of the relation		1975-1980
	<i>ISO 8601</i>	1975/1980
<i>Eighth Relation</i>		
5.3.1 Name / identifier of the related entity	<i>Authorized form of name</i>	Australia. National Aboriginal Education Committee
5.3.2 Category of relationship		Associative
		Committee member

5.3.3 Description of relationship	<i>Narrative</i>	The National Aboriginal Education Committee was set up to provide advice to the Minister of Education and the Department of Education on Aboriginal views on the educational needs of Aboriginal people, and to monitor existing policies and programs. Mabo became involved in this Committee through his work for the Black Community School, and was a Committee member between 1975 and 1978.
5.3.4 Dates of the relation		1975-1978
	<i>ISO 8601</i>	1975/1978
<i>Ninth Relation</i>		
5.3.1 Name / identifier of the related entity	<i>Authorized form of name</i>	Australia. Commonwealth Employment Service. Aboriginal Employment and Training Branch
5.3.2 Category of relationship		Associative
5.3.3 Description of relationship		Employee
	<i>Title</i>	Assistant Vocational Officer
5.3.4 Dates of the relationship		1978-1981
	<i>ISO 8601</i>	1978/1981
<i>Tenth Relation</i>		
5.3.1 Name / identifier of the related entity	<i>Authorized form of name</i>	ABIS Community Cooperative Society Ltd (Townsville, Qld)
5.3.2 Category of relationship		Associative
5.3.3 Description of relationship	<i>Title</i>	Director
	<i>Narrative</i>	The ABIS Community Cooperative Society was a Townsville-based Aboriginal and Islander cooperative housing association.
5.3.4 Dates of the relationship		1986-1987
	<i>ISO 8601</i>	1986/1987
<i>Eleventh Relation</i>		
5.3.1 Name / identifier of the related entity	<i>Authorized form of name</i>	Moomba Festival (Melbourne, Vic.)
5.3.2 Category of relationship		Associative
5.3.3 Description of relationship		Employee
	<i>Title</i>	Assistant Director, Aboriginal Arts
	<i>Narrative</i>	During 1986-87, Mabo participated in the Communication and Arts Management Scheme run by the Aboriginal Training and Cultural Institute. Through this Scheme he was appointed Assistant Director, Aboriginal Arts, Melbourne Moomba Festival. Mabo claimed that his efforts ensured the first-ever Aboriginal involvement in the Moomba Festival.
5.3.4 Dates of the relationship		1986-1987
	<i>ISO 8601</i>	1986/1987
<i>Twelfth Relation</i>		
5.3.1 Name/identifier of the related entity	<i>Authorized form of name</i>	Festival of Pacific Arts (5th: 1988: Townsville, Qld)
5.3.2 Category of relationship		Associative
	<i>Title</i>	Liaison Officer

5.3.3 Description of relationship	<i>Narrative</i>	The 5th Festival of Pacific Arts, which took place in Townsville in 1988, was the first to be held in Australia. The Festival of Pacific Arts occurs every four years and is organised under the auspices of the South Pacific Commission. The 1988 Festival received funding from the Australian Government through the Department of Arts, Heritage and the Environment. Mabo was employed by the Department of Aboriginal Affairs as Community Arts Liaison Officer for the Festival, 1987-88.
5.3.4 Dates of the relationship		1987-1988
	<i>ISO 8601</i>	1987/1988
<i>Thirteenth Relation</i>		
5.3.1 Name / identifier of the related entity	<i>Authorized form of name</i>	Australia. Dept of Aboriginal Affairs
5.3.2 Category of relationship		Associative
5.3.3 Description of relationship		Employee
	<i>Title</i>	Liaison Officer, 5th Festival of Pacific Arts, Townsville, Qld
5.3.4 Dates of the relationship		1987-1988
	<i>ISO 8601</i>	1987/1988
<i>Fourteenth Relation</i>		
5.3.1 Name / identifier of the related entity	<i>Authorized form of name</i>	Magani Malu Kes
5.3.2 Category of relationship		Associative
5.3.3 Description of relationship	<i>Title</i>	Vice-Chairman
	<i>Narrative</i>	Magani Malu Kes is the name for the Torres Strait Islands in the language of the Torres Strait. The organisation Magani Malu Kes was an organisation for Torres Strait Islanders, which Mabo had incorporated as a public company in 1987. Of major concern to Magani Malu Kes was the way in which Islander interests appeared to be marginalised by those of mainland Aborigines when indigenous issues were considered by governments. As a consequence, Magani Malu Kes advocated Torres Strait Islander independence from Australia.
5.3.4 Dates of the relationship		1987-1988
	<i>ISO 8601</i>	1987/1988
<i>Fifteenth Relation</i>		
5.3.1 Name / identifier of the related entity	<i>Authorized form of name</i>	Australia. High Court
5.3.2 Category of relationship		Associative
5.3.3 Description of relationship	<i>Title</i>	Plaintiff
	<i>Narrative</i>	In 1981, at a conference on indigenous land rights in Townsville, a decision was made to pursue a native land title claim for the people of the Murray Islands in the High Court of Australia. In 1982, Mabo and four other Islander plaintiffs instituted proceedings against the State of Queensland, claiming that their islands had been continuously inhabited and exclusively possessed by their people who lived in permanent settled communities. They acknowledged that the British Crown became sovereign of the islands upon annexation, but claimed continuous enjoyment of their land rights which had not been validly extinguished by the sovereign through the granting of freehold title or land leases to others. The Queensland Government attempted to defeat the claim with the passage of the <i>Queensland Coast Islands Declaratory Act</i> 1985. In 1988, the High Court ruled this Act contrary to the Commonwealth <i>Racial Discrimination Act</i> 1975. In May 1989, the High Court remitted the land claim to the Queensland Supreme Court for hearing and determination of all issues of fact. In November 1990, Justice

		Moynihan of the Supreme Court delivered the Court's determination of the issues of fact. The case was argued for four days before the High Court in May 1991. The final decision was handed down in favour of Mabo on 3 June 1992. This decision overturned the 204- year-old legal doctrine of <i>terra nullius</i> , which held that the lands of the Australian continent were 'practically unoccupied' at the time of the proclamation of British sovereignty.
5.3.4 Dates of the relationship		1985-1992
	ISO 8601	1985/1992

Sixteenth Relation		
5.3.1 Name / identifier of the related entity	Authorized form of name	Murray Island Community Council
5.3.2 Category of relationship		Associative
5.3.3 Description of relationship	Narrative	During the late 1980s Mabo attempted to gain election to the Murray Island [Mer] Community Council. However, because he had not lived on Mer since the late 1950s, his residential status was questioned and it was ruled that he was not eligible to nominate.
5.3.4 Dates of the relationship		1985-1991
	ISO 8601	1985/1991
Seventeenth Relation		
5.3.1 Name / identifier of the related entity	Authorized form of name	Australian Institute of Aboriginal Studies. Education Advisory Committee
	Successor	Australian Institute of Aboriginal and Torres Strait Islander Studies. Education Advisory Committee
5.3.2 Category of relationship		Associative
5.3.3 Description of relationship	Narrative	Located in Canberra, the Australian Institute of Aboriginal Studies (later the Australian Institute of Aboriginal and Torres Strait Islander Studies) promotes and supports research into the cultures (both traditional and contemporary), languages, histories, and contemporary needs of Australia's indigenous communities. Mabo first became associated with the Institute in 1978 when, as Director of the Black Community School, he was appointed to its Education Advisory Committee.
5.3.4 Dates of the relationship		1978-1989
	ISO 8601	1978/1989
5.4 CONTROL AREA		
5.4.1 Authority record identifier		AU 93-435878
5.4.2 Institution identifiers		National Library of Australia
	ILL Code	AU NLA

5.4.3 Rules and/or conventions		<ul style="list-style-type: none"> - ISAAR (CPF) – <i>International Standard Archival Authority Record For Corporate Bodies, Persons and Families</i>, Draft 2nd ed., Madrid: International Council on Archives, 12-15 June 2002. - <i>Anglo American Cataloguing Rules</i> 2nd rev. ed., Chicago, 1998. - ISO 8601 - <i>Data elements and interchange formats –Information interchange—Representation of dates and times</i>, 2nd ed., Geneva: International Standards Organization, 2000. - ISO 3166 - <i>Codes for the representation of names of countries</i>, Geneva: International Standards Organization, 1997. - ISO 15511 - <i>Information and documentation - International Standard Identifier for Libraries and Related Organisations (ISIL)</i>, Geneva: International Standards Organization, 2000. - ISO 639-2 - <i>Codes for the representation of names of languages - Part 2: Alpha-3 Code</i>, Geneva: International Standards Organization, 1998. - ISO 15924 - <i>Codes for the representation of names of scripts</i>, Geneva: International Standards Organization, 2001.
5.4.4 Status		Revised
5.4.5 Level of detail		Full
5.4.6 Dates of creation and revision	ISO 8601	1993-05-12; revised 2002-10-28
5.4.7 Languages and scripts		English
	ISO 639-1	en
	ISO 15024	latn
5.4.9 Maintenance notes	Creator of authority record	Adrian Cunningham
6. RELATING CORPORATE BODIES, PERSONS AND FAMILIES TO ARCHIVAL MATERIALS AND OTHER RESOURCES		
<i>First Related ReSource</i>		
6.1 Identifier and title of related reSource	Title	Papers of Eddie Koiki Mabo
	Unique Identifier	AU NLA MS 8822
6.2 Type of related reSource		Personal papers
6.3 Nature of relationship		Creator
6.4 Dates of related resources and/or relationships		1943, 1959-1992 (bulk: 1972-1992)

<i>Second Related ReSource</i>		
6.1 Identifier and title of related source	Title	Guide to the papers of Edward Koiki Mabo in the National Library of Australia
	Unique ID	http://www.nla.gov.au/ms/findaids/8822.html#sd
6.2 Type of related reSource		Finding aid
6.3 Nature of relationship		Subject
6.4 Dates of related resources and/or relationships		1995
<i>Third Related ReSource</i>		
6.1 Identifier and title of related source	Title	Papers of Edward Koiki Mabo [microfilm]
	Unique Identifier	AU NLA PRU Mfm G 27539-27549 (copying master : Manuscripts) Mfm G 27539-27549 PRU Mfm G 27539-27549 (first generation master : coldstore) Mfm G 27623

6.2 Type of related reSource		Microfilm copy of personal papers
6.3 Nature of relationship		Creator
6.4 Dates of related resources and/or relationships		1996
<i>Fourth Related ReSource</i>		
6.1 Identifier and title of related source	<i>Title</i>	Records of Brian Keon-Cohen
	<i>Unique Identifier</i>	AU NLA MS 9518
6.2 Type of related reSource		Archival materials
6.3 Nature of relationship		Subject. Records relating to the Mabo case. Mabo Litigation Records emanating from both the Supreme Court of Queensland and the High Court of Australia. They comprise a Statement of Facts by the plaintiffs, wills, land transactions, court transcripts, exhibits, pleadings, applications, witness statements, submissions, correspondence, memoranda and research material. Keon-Cohen, with the assistance of an archivist engaged at the Library's expense, arranged the items into volumes in broad chronological order. An index was compiled by the archivist.
6.4 Dates of related resources and/or relationships		1982-1992
<i>Fifth Related ReSource</i>		
6.1 Identifier and title of related source	<i>Title</i>	<i>Edward Koiki Mabo : his life and struggle for land rights/</i> by Noel Loos
	<i>Unique Identifier</i>	ISBN 0702229059
6.2 Type of related reSource		Monograph biography
6.3 Nature of relationship		Subject
6.4 Dates of related resources and/or relationships		1996
<i>Sixth Related ReSource</i>		
6.1 Identifier and title of related source	<i>Title</i>	<i>Mabo: Life of an Island Man</i>
	<i>Unique Identifier</i>	
6.2 Type of related reSource		Videorecording
6.3 Nature of relationship		Subject
6.4 Dates of related resources and/or relationships		1996

Example 10 - Family description

Language of description: English (United Kingdom)

5.1 IDENTITY AREA		
5.1.1 Type of entity		Family
5.1.2 Authorized form of name		Noel family, Earls of Gainsborough
5.1.5 Other forms of name		Noel family, Barons Noel Noel family, Barons Barham Noel family, Viscounts Campden Noel family, baronets, of Barham Court
5.2 DESCRIPTION AREA		
5.2.1 Dates of existence		12 th – 20 th century

5.2.2 History		<p>The Noel family was established in Staffordshire in the Middle Ages: Andrew Noel (d1563), third son of James Noel of Hidcote (Staffordshire) acquired property in Rutland and Leicestershire and founded the branch of the Noel family from which the Earls of Gainsborough descended. Estates in other counties (mainly Gloucestershire, Hampshire, Kent and Middlesex) were subsequently acquired through purchase, inheritance and marriage.</p> <p>The family's principal seat was Exton Hall (Rutland). After Exton Hall was severely damaged by fire in 1810 Barham Court near Maidstone (Kent) became the main residence until the sale of the Kent estate in 1845-6 but a new house at Exton was completed in the early 1850s. Campden House in Chipping Campden (Gloucestershire), inherited from Baptist Hicks, Viscount Campden (d 1629), was destroyed in the Civil War while Campden House (Kensington) was sold in 1708. The family of Gerard Noel Edwards, who inherited the Noel family estates in 1798 and took the surname Noel, had resided at Welham Grove in Welham (Leicestershire) but this was sold by 1840.</p> <p>Andrew Noel (d1563) acquired Old Dalby (Leicestershire, sold 1617) and Brooke (Rutland). His grandson Sir Edward Noel (d 1643) bought the former Harington family estate of Ridlington (Rutland) and was created Baron Noel in 1617. He married a co-heir of Baptist Hicks (Viscount Campden, d 1629), through whom came the property in Rutland (Exton and Whitwell), Gloucestershire (Chipping Campden), Middlesex (Hampstead, sold 1707) and Campden House (Kensington, sold 1708), and succeeded his father-in-law as second Viscount Campden. Valle Crucis (Denbighshire) was acquired through the marriage of the third Viscount Campden (1612-82) and Hester, daughter of the second Baron Wotton, but was sold in 1663 to Sir John Wynn, fifth Bt. The fourth Viscount Campden (1641-89, created Earl of Gainsborough 1682) married in 1661 Elizabeth Wriothesley, elder daughter of the fourth Earl of Southampton, through whom came the Titchfield (Hampshire) estate, but on the second Earl of Gainsborough's death in 1690 this estate passed to his daughters (who married respectively the first Duke of Portland and the second Duke of Beaufort). The Rutland and Gloucestershire estates, with the earldom of Gainsborough, however, were inherited by the second Earl's cousin Baptist Noel of North Luffenham and Cottesmore (both Rutland).</p> <p>Property at Walcot (Northamptonshire), Castle Bytham (Lincolnshire) and Kinnoulton (Nottinghamshire) was inherited by the sixth Earl of Gainsborough from a cousin, Thomas Noel of Walcot (d 1788). On the sixth Earl's death in 1798 his estates were divided, the Walcot properties passing to CH Nevile, who took the name Noel, and the Rutland and Gloucestershire estates passing to Gerard Noel Edwards (1759-1838), later Sir GN Noel, second Bt, son of GA Edwards (d 1773) of Welham Grove (Leicestershire) by Jane Noel (d 1811), sister of the fifth and sixth Earls. GN Edwards had inherited from his father various properties in Leicestershire (Welham, etc), London and Middlesex (Islington, Kensington, etc), Kent (Dartford, etc) and Ireland, but these were all sold between 1805 and 1840. GN Edwards had married in 1780 Diana, daughter of Charles Middleton (1726- 1813), first Baron Barham, who succeeded her father in the barony and the Barham Court estate near Maidstone (Kent). On her death in 1823 she was in turn succeeded in the Barham Court estate and peerage title by her son Charles Noel (d 1866), who inherited the Rutland (Exton, etc) and Gloucestershire (Chipping Campden) estates of his father in 1838 and was created Earl of Gainsborough in 1841.</p> <p>The remaining Kent property was, however, sold by 1845-46.</p>
5.2.3 Places		<p>Estates in 1883: Rutland 15,076 acres, Gloucestershire 3,170 acres, Leicestershire 159 acres, Lincolnshire 89 acres, Warwickshire 68 acres, Northamptonshire 6 acres; total 18,568 acres worth £28,991 a year.</p>

5.2.5 Functions, occupations and activities		Estate ownership; social, political and cultural role typical of the landed aristocracy in England. The first Viscount Campden amassed a large fortune in trade in London and purchased extensive estates, including Exton (Rutland) and Chipping Campden (Gloucestershire). The Barham Court (Kent) estate was the acquisition of the first Baron Barham, a successful admiral and naval administrator (First Lord of the Admiralty 1805).
5.2.7 Genealogy		Sir Edward Noel (d 1643) married Julian, daughter and co-heir of Baptists Hicks (d 1629), Viscount Campden, and succeeded to the viscounty of Campden and a portion of his father-in-law's estates. The third Viscount Campden (1612-82) married Hester Wotton, daughter of the second Baron Wotton. The fourth Viscount Campden (1641-89, created Earl of Gainsborough 1682) married Elizabeth Wriothesley, elder daughter of the fourth Earl of Southampton. Jane Noel (d 1811), sister of the fifth and sixth Earls of Gainsborough, married Gerard Anne Edwards of Welham Grove (Leicestershire) and had issue Gerard Noel Edwards (1759-1838). He married in 1780 Diana Middleton (1762- 1823) <i>suo jure</i> Baroness Barham), daughter of Charles Middleton (1726-1813), created first Baronet of Barham Court (Kent) in 1781 and first Baron Barham in 1805. GN Edwards assumed the surname Noel in 1798 on inheriting the sixth Earl of Gainsborough's Rutland and Gloucestershire estates (though not the Earl's honours, which were extinguished); and he later inherited his father-in-law's baronetcy. His eldest son John Noel (1781-1866) succeeded to the estates of his mother and his father, to his mother's barony and his father's baronetcy, and was created Viscount Campden and Earl of Gainsborough in 1841.
5.3 RELATIONSHIPS AREA		
<i>First Relation</i>		
5.3.1 Name / identifier of the related entity	<i>Authorized form of name</i>	Harington family, Barons Harington Family
	<i>Identifier</i>	GB/NNAF/F10219
5.3.2 Category of relationship		Family
5.3.3 Description of relationship		Predecessor in the Ridlington (Rutland) estate
5.3.4 Dates of the relationship		Early 17 th century
<i>Second Relation</i>		
5.3.1 Name / identifier of the related entity	<i>Authorized form of name</i>	Wotton family, Barons Wotton
	<i>Identifier</i>	GB/NNAF/F10218
5.3.2 Category of relationship		Family
5.3.3 Description of relationship		Third Viscount Campden married Hester, daughter of second Baron Wotton
5.3.4 Dates of the relationship		Mid 17 th century
<i>Third Relation</i>		
5.3.1 Name / identifier of the related entity	<i>Authorized form of name</i>	Bentinck, Cavendish- family, Dukes of Portland
	<i>Identifier</i>	GB/NNAF/F9541
5.3.2 Category of relationship		Family
5.3.3 Description of relationship		A daughter of second Earl of Gainsborough married the first Duke of Portland
5.3.4 Dates of the relationship		Late 17 th century
<i>Fourth Relation</i>		
5.3.1 Name / identifier of the related entity	<i>Authorized form of name</i>	Somerset family, Dukes of Beaufort
	<i>Identifier</i>	GB/NNAF/F3483
5.3.2 Category of relationship		Family

5.3.3 Description of relationship		The second Duke of Beaufort married Rachel daughter and coheir of the second Earl of Gainsborough in 1706
5.3.4 Dates of the relationship		1706
<i>Fifth Relation</i>		
5.3.1 Name / identifier of the related entity	<i>Authorized form of name</i>	Wriothesley family, Earls of Southampton
	<i>Identifier</i>	GB/NNAF/F2938
5.3.2 Category of relationship		Family
5.3.3 Description of relationship		Elizabeth, daughter of the fourth Earl of Southampton married the first Earl of Gainsborough in 1661
5.3.4 Dates of the relationship		Late 17 th century

66

<i>Sixth Relation</i>		
5.3.1 Name / identifier of the related entity	<i>Authorized form of name</i>	Noel family of Walcot
	<i>Identifier</i>	GB/NNAF/F10217
5.3.2 Category of relationship		Family
5.3.3 Description of relationship		The sixth Earl of Gainsborough inherited the Walcot (Northamptonshire) estates of his cousin Thomas Noel in 1788
5.3.4 Dates of the relationship		1788
<i>Seventh Relation</i>		
5.3.1 Name / identifier of the related entity	<i>Authorized form of name</i>	Edwards family of Welham
	<i>Identifier</i>	GB/NNAF/F7310
5.3.2 Category of relationship		Family
5.3.3 Description of relationship		GN Edwards inherited the Noel estates in 1798 and took the surname Noel
5.3.4 Dates of the relation		1798
5.4 CONTROL AREA		
5.4.1 Authority record identifier		GB/NNAF/F10216
5.4.2 Institution identifiers		Historical Manuscripts Commission
5.4.3 Rules and/or conventions		National Council on Archives <i>Rules for the Construction of Personal Place and Corporate Names</i> , 1997
5.4.4 Status		Finalised
5.4.5 Level of detail		Full
5.4.6 Dates of creation and revision		30 November 2000
5.4.7 Languages and scripts		English
5.4.8 Sources		HMC, <i>Principal Family and Estate Collections: Family Names L-W</i> , 1999 <i>Complete Peerage</i> , 1936 <i>Burkes Peerage</i> , 19q70 <i>Complete Baronetage</i> , vol 5, 1906

534

6. RELATING CORPORATE BODIES, PERSONS AND FAMILIES TO ARCHIVAL MATERIALS AND OTHER RESOURCES		
<i>First Related ReSource</i>		
6.1 Identifier and title of related reSource	<i>Title</i>	Family and estate papers
	<i>Unique Identifier</i>	GB 0056 DE 3214
6.2 Type of related reSource		Archival materials
6.3 Nature of relationship		Creator
6.4 Dates of related resources and/or relationships		12 th -20 th cent
<i>Second Related ReSource</i>		
6.1 Identifier and title of related reSource	<i>Title</i>	Rutland estate sales papers
	<i>Unique Identifier</i>	GB 0056 DE 3177/36-44
6.2 Type of related reSource		Archival materials
6.3 Nature of relationship		Creator
6.4 Dates of related resources and/or relationships		1925-26
<i>Third Related ReSource</i>		
6.1 Identifier and title of related reSource	<i>Title</i>	Deeds, family financial and trust papers
	<i>Unique Identifier</i>	GB 0056 DE 2459
6.2 Type of related reSource		Archival materials
6.3 Nature of relationship		Creator
6.4 Dates of related resources and/or relationships		17 th -19 th cent
<i>Fourth Related ReSource</i>		
6.1 Identifier and title of related reSource	<i>Title</i>	Pickwell (Leicestershire) estate maps
	<i>Unique Identifier</i>	GB 0056 89-91/30
6.2 Type of related reSource		Archival materials
6.3 Nature of relationship		Creator
6.4 Dates of related resources and/or relationships		1616, 1736

67

<i>Fifth Related ReSource</i>		
6.1 Identifier and title of related reSource	<i>Title</i>	Deeds, family and estate papers
	<i>Unique Identifier</i>	GB 0056 DE 1797
6.2 Type of related reSource		Archival materials
6.3 Nature of relationship		Creator
6.4 Dates of related resources and/or relationships		13 th -18 th cent
<i>Sixth Related ReSource</i>		
6.1 Identifier and title of related reSource	<i>Title</i>	Welham (Leicestershire) deeds and estate papers
	<i>Unique Identifier</i>	GB 0056 81'30
6.2 Type of related reSource		Archival materials

535

6.3 Nature of relationship		Creator
6.4 Dates of related resources and/or relationships		1745-1838
<i>Seventh Related ReSource</i>		
6.1 Identifier and title of related reSource	<i>Title</i>	Chipping Campden (Gloucestershire) deeds and papers
	<i>Unique Identifier</i>	GB 0056 DE 3214
6.2 Type of related reSource		Archival materials
6.3 Nature of relationship		Creator
6.4 Dates of related resources and/or relationships		15 th -20 th cent
<i>Eighth Related ReSource</i>		
6.1 Identifier and title of related reSource	<i>Title</i>	Chipping Campden (Gloucestershire) deeds and papers
	<i>Unique Identifier</i>	GB 0040 D329
6.2 Type of related reSource		Archival materials
6.3 Nature of relationship		Creator
6.4 Dates of related resources and/or relationships		1707-1881
<i>Ninth Related ReSource</i>		
6.1 Identifier and title of related reSource	<i>Title</i>	Titchfield (Hampshire) deeds and estate papers
	<i>Unique Identifier</i>	GB 0041 5M53
6.2 Type of related reSource		Archival materials
6.3 Nature of relationship		Creator
6.4 Dates of related resources and/or relationships		13 th -18 th cent
<i>Tenth Related ReSource</i>		
6.1 Identifier and title of related source	<i>Title</i>	A-E Noel and Edwards family corresp and papers 18 th -20 th cent
	<i>Unique Identifier</i>	GB 800819
6.2 Type of related reSource		Archival materials
6.3 Nature of relationship		Creator
6.4 Dates of related resources and/or relationships		18 th -20 th cent
<i>Eleventh Related ReSource</i>		
6.1 Identifier and title of related reSource	<i>Title</i>	Noel family seal
	<i>Unique Identifier</i>	GB 0066, E 40/12531
6.2 Type of related reSource		Attached seal, Andrew Noel
6.3 Nature of relationship		Owner
6.4 Dates of related resources and/or relationships		1551-1552

Appendix 6 Describing Archives: A Content Standard (DACS)

The following select table of access point (and other relative) standards is taken from Describing Archives: A Content Standard – DACS 2020 (Society of American Archivists 2020)

An archival authority record identifies and describes a personal, family, or corporate entity associated with a body of archival materials, documents relationships between records creators, the records created by them, and/or other resources about them; and may control the creation and use of access points in archival descriptions. The International Standard Archival Authority Record for Corporate Bodies, Persons and Families (ISAAR[CPF]) organizes the types of information found in an archival authority record into four areas:

- Identity Area: the authoritative form of the name of the entity as established by cataloguing rules such as those found in AACR2 or RDA, along with references to any variant forms of that name by which researchers might know that entity
- Description Area: a description of the history and activities of the entity that are pertinent to the records with which it is associated, written in accordance with the rules in Chapter 11
- Relationships Area: references to related persons, families, and corporate bodies

- Control Area: management information regarding the creation and status of the record

Although archival authority records are similar to library authority records in that they both support the creation of standardized access points in descriptions, archival authority records support a much wider set of requirements than library authority records do and usually contain detailed information about records creators and the context of record creation. (Society of American Archivists 2020, 85)

Authority information may be used in a variety of ways. It can provide access to archival materials based on descriptions of records creators or the context of records creation that are linked to descriptions of physically dispersed records. It can provide users an understanding of the context underlying the creation and use of archival materials so they can better interpret their meaning and significance. It can help users identify records creators by providing descriptions of relationships between different entities, particularly in cases of administrative changes within corporate bodies or personal changes in families and individuals. Finally, standardized authority information allows for the exchange of descriptions of individuals, families, and corporate bodies between institutions, systems, and networks and across national and linguistic boundaries. (Society of American Archivists 2020, 86)

An authority record with the minimum number of DACS elements includes:
<ul style="list-style-type: none"> • Authorized form of name (see 10.1)
<ul style="list-style-type: none"> • Type of entity (see 10.2)
<ul style="list-style-type: none"> • Dates of existence (see 11.1)
<ul style="list-style-type: none"> • Authority record identifier (see 13.2)
Take the information from any reliable source. [and] create an authority record for each person, family, or corporate body associated with the creation of archival materials as specified in the rule (DACS 87)
Record the name of the entity being described in the authority record in accordance with standardized vocabularies (e.g., LCNAF or with rules for formulating standardized names such as those found in AACR2, RDA, or ISAAR(CPF). Name entry may include dates, place, jurisdiction, occupation, epithet, or other qualifiers.
<ul style="list-style-type: none"> • Haworth, Kent MacLean, 1946-
<ul style="list-style-type: none"> • Stibbe, Hugo L. P.

<ul style="list-style-type: none"> • Cadell, T. (Thomas), 1742-1802 (DACS 88)
Record all other names or forms of name(s) that might reasonably be sought by a user but were not chosen as the authorized form of name. (DACS 90)
11.1.1 Record dates associated with the entity being described. Record dates in terms of the calendar preferred by the agency creating the data. Record dates in the following formats:
<ul style="list-style-type: none"> • Record exact dates in [year] [month] [day] format.
<ul style="list-style-type: none"> • Indicate a probable date by adding a question mark following the year.
<ul style="list-style-type: none"> • If the year is uncertain but known to be either one of two years, record the date in the form [year] or [year].
<ul style="list-style-type: none"> • If the year can only be approximated, record the date in the form approximately [year].
11.1.2 For a person, record his or her date of birth and/or date of death. Where exact dates are not known, record approximate dates.
1884 May 8 (date of birth)

1796? (date of birth)
1501 or 1507 (date of birth)
1826 July 4 (date of death)
approximately 1945 January (date of death)
1972
1742 November 12-1802 December 27
11.1.3 For a person, if both the date of birth and date of death are unknown, record floruit (period of activity) dates. If specific years of activity cannot be established, record the century or centuries in which the person was active.
1841-1874 (active)
12th century (active) (DACS 93)
For families, record significant dates associated with the family such as establishment dates or floruit dates. If specific years cannot be established, record the century or centuries in which the family was active.

1802 (date of establishment)
1945 (date of termination)
ninth century (end date of activity) (DACS 94)
<p>Internal Structure/Genealogy.</p> <p>11.7.1 Record in narrative form the internal structure of the entity being described.</p> <p>Wherever possible, devise dates as an integral component of the narrative description.</p> <p>(DACS 100)</p>
<p>Example Description of the Person, Family, or Corporate Body Area of an Archival</p> <p>Authority Record</p>
Dates of Existence (11.1.2): 1742 November 12-1802 December 27
Historical Summary (11.2.1):

Thomas Cadell was born in Bristol on 12 November 1742 but spent most of his life in London. When Cadell was fifteen, his father sent him to be an apprentice to Andrew Millar

(1707-1 of Samuel Johnson's Dictionary. After seven years, Cadell became a partner in the business 768), a well-regarded publisher and bookseller who had supported the publication and finally took it over when Millar retired in 1767. His clients and friends were among the most influential literary and intellectual figures of the eighteenth century and included

Fanny Burney (1752-1840), Robert Burns (1759-1796), David Hume (1711-1776), Samuel

Johnson (1709-1784), Hannah More (1745-1833), Adam Smith (1723-1790), and Tobias

Smollett (1721-1771). When Cadell retired in 1793, he gave his business to his son,

Thomas Cadell (1773-1836) and his former assistant, William Davies (d. 1820). Before his death from an asthma attack in 1802, he enjoyed an active retirement, fulfilling many charitable and public positions, including governor of the Foundling Hospital and sheriff in the Walbrook ward of London.

Places (11.3.2):

Born: Bristol (England)
Lived: London (England)
Functions, Occupations, Activities (11.5.2):
Booksellers
Publishers
Stationers (DACS 101)
<p>Commentary: In describing the parties that created, assembled, accumulated, and/or maintained and used archival records, it will be useful to identify related persons, families, and organizations. They may be connected in a variety of ways, such as members of families, hierarchical relationships between parts of organizations, chronological (i.e., predecessor/successor) relationships between organizations or parts of organizations, or offices held by a person within an organization. Related names might also be used within a descriptive system as alternative access points to descriptions of archival records or as links to other authority records. Record the authorized names and any relevant unique identifiers, including the authority record identifier, of corporate bodies, persons, or</p>

families that have a significant relationship with the entity named in the authority record.

(DACS 102)

Provide a repository code for the institution creating the authority record. Use the repository codes assigned by the national organization responsible for assigning and maintaining repository identifiers or appropriate international repository identifiers.

(DACS 107)

Record a unique identifier for the authority record. The number may be assigned locally or be based upon an identifier from a regional or national database such as the Library of Congress Authorities. (DACS 108)

Record relevant information about Sources consulted in establishing or revising the authority record. Establish a consistent policy regarding the content, form, and placement of citation of Sources. (DACS 114)

Record the name(s) of the person(s) who prepared or revised the authority record and any other information pertinent to its creation or maintenance. (DACS 115)

Appendix 7 The Project Seven Report

The JupyterBook pdf will be inserted here. This page is intentionally blank.

Project Seven

Contents

- P7 Chapter 1 - The Human Data Digital Toolkit (HDDT)
- 1.1 The data
- 1.2 The HDDT model:
- 1.3 Project Seven Questions
- 1.4 Contributing datasets
- 1.5 Dataset variability
- 1.6 Data pipeline
- 1.7 Model output - Data visualisation criteria
- 1.8 Integrated HTTD SQLite database - 3095 persons
- 1.9 HDDT design architecture
- 1.10 Project containers must contain all required resources
- 1.11 GitHub
- 1.12 SQLite recommendations
- 1.13 VSC recommendations
- 1.14 VSC Version control interface - local to online Git Repo's
- 1.15 DBeaver recommendations
- 1.16 Jupyter Notebook recommendations
- 1.17 Gephi recommendations
- 1.18 Excel recommendations
- 1.19 References
- 1.20 600 Quakers amongst 3000 activists for 40 years 1830-1870
- P7 Chapter 2 The CEDA Members
- 2.1 Introduction
- 2.2 The Entity Relationship Diagram

- 2.3 ERD statistics
- 2.4 The structure and dimensions of all SQL tables
- 2.5 The CEDA
- 2.6 All CEDA members' relationships visualisation
- 2.7 All persons are members of at least one CEDA society
- 2.8 Memberships in each CEDA table
- 2.9 CEDA members were also members of 68 clubs
- 2.10 CEDA members are identified with 83 locations
- 2.11 top 10 locations
- 2.12 Some persons are associated with multiple locations
- 2.13 CEDA members occupations
- 2.14 the top 10 Occupations
- 2.15 Some members have more than one occupation
- 2.16 Society memberships
- 2.17 Top 10 Society memberships
- 2.18 Some CEDA members are members of multiple societies
- 2.19 All Quaker relationships
- 2.20 Quaker family relationships
- 2.21 Quaker members of the CEDA
- 2.22 - All SQL relatable tables rendered in Gephi format
- P7 Chapter 3 HDDT Using SQLite database standard 'views'
- 3.1 Introduction and explanation
- 3.2 To make a new project
- 3.3 Entity Relationship Diagram
- 3.4 Person table (3094 records)
- 3.5 Person Names (3094 records)
- 3.6 Persons with attributes (Names file) (3094 records)
- 3.7 All Names (Nodes) (3608 records)
- 3.8 All bipartite Names (514 records)
- 3.9 All Names (Nodes) as Tuples (9989 records)

- 3.10 Religion tuples (592 records)
- 3.11 Location tuples (2061 records)
- 3.12 Occupation tuples (1883 records)
- 3.13 Society tuples (1238 records)
- 3.14 Club tuples (323 records)
- 3.15 CEDA tuples (3892 records)
- 3.16 Quaker Committee on the Aborigines (QCA)
- 3.17 Aborigines Protection Society (APS)
- 3.18 Ethnological Society of London (ESL)
- 3.19 Anthropological Society of London (ASL)
- 3.20 Anthropological Institute (AI)
- 3.21 CEDA Name with attributes (3892 records)
- 3.22 CEDA tuples with attributes (3892 records)
- 3.23 Quakers (592 records)
- 3.24 Quaker family relationships (2086 records)
- 3.25 Quaker immediate relationships (246 records)
- 3.26 Quakers close relationships (519 records)
- 3.27 Quaker distant relationships (1321 records)
- 3.28 Quaker CEDA membership (tuples) (643 records)
- 3.29 Quakers in the QCA
- 3.30 Quakers in the APS
- 3.31 Quakers in the ESL
- 3.32 Quakers in the ASL
- 3.33 Quakers in the AI
- P7 Chapter 4 Case Study 1 The Centres for the Emergence of the Discipline of Anthropology (CEDA)
- 4.1 HDDT Visualisations - CEDA bigraph
- 4.2 The CEDA 1830 - 1870
- 4.3 The Quaker Committee on the Aborigines (QCA) 1837 -1846
- 4.4 CQA Joiners each year

- 4.5 CQA Leavers each year
- 4.6 Duration in the CEDA
- 4.7 The Aborigines Protection Society (APS) 1837 -1919
- 4.8 APS joiners in each year
- 4.9 APS leavers in each year
- 4.10 Quakers joining the APS in each year
- 4.11 Quakers leaving the APS in each year
- 4.12 The Ethnological Society of London (ESL) 1843 - 1871
- 4.13 ESL joiners in each year
- 4.14 ESL leavers in each year
- 4.15 ESL Quaker joiners in each year
- 4.16 ESL Quaker leavers in each year
- 4.17 The Anthropological Society of London (ASL) 1863 - 1871
- 4.18 ASL joiners in each year
- 4.19 ASL leavers in each year
- 4.20 ASL Quaker joiners in each year
- 4.21 ASL Quaker leavers in each year
- 4.22 Anthropological Institute (AI) 1843 - 1871
- 4.23 AI joiners in each year
- 4.24 AI leavers in each year
- 4.25 AI Quaker joiners in each year
- 4.26 AI Quaker leavers in each year
- 4.27 Duration of AI Quaker memberships
- 4.28 generate Gexf output file of all CEDA data for Gephi
- P7 Chapter 5a Thomas Hodgkin MD's networks - Part one
- 5.1 Protecting the Empire's Humanity: Thomas Hodgkin and British Colonial Activism 1830 - 1870 (Zoë Laidlaw 2021)
- 5.2 GitHub
- 5.3 Call up the python packages needed to perform the analysis
- 5.4 Call up the csv files from the SQL db and prepare data for Gephi

- 5.5 Introduction to the exercise - Part One
- 5.6 The 3094 members of the CEDA before the exercise
- 5.7 Mods to db to facilitate this exercise - ZOE and WEL
- 5.8 Assess persons in the index to PEH who are members of the CEDA.
- 5.9 CEDA members compared to non-CEDA others in PEH index
- 5.10 Data verification
- 5.11 Generate gexf file for Gephi visualisation
- 5.12 Visual analysis of the exercise
- 5.13 The CEDA social network including ZOE and WEL
- 5.14 Zooming in to show the network in detail
- 5.15 Other groupings emerge
- 5.16 Quaker roles emerge in detail
- 5.17 Conclusions
- 5.18 Modifications to the database for Part Two
- 5.19 Github upload
- P7 Chapter 5b Thomas Hodgkin's MD networks - Part two
- *Protecting the Empire's Humanity* (PEH): Thomas Hodgkin and British Colonial Activism 1830 - 1870 (Zoë Laidlaw 2021)
- 5.20 Preparation
- 5.21 GitHub
- 5.22 Call up the Python packages needed to perform the analysis
- 5.23 call up the csv files and prepare data for Gephi visualisation
- 5.24 Introduction to the exercise - Part Two
- 5.25 Data verification
- 5.26 Person table after modification
- 5.27 Persons who are members of the new HOD CEDA
- 5.28 Generate gexf file for Gephi visualisation
- 5.29 Visual analysis of the exercise
- 5.30 The CEDA political network with HOD added
- 5.31 The CEDA political network with HOD added in detail

- 5.32 GitHub upload
- P7 Chapter 6 Case Study 2 589 Quakers and their family relationships
- 6.1 Import resources
- 6.2 List out all Quakers in the database
- 6.3 List out all the Quaker family relationships
- 6.4 All Quaker family relationships
- 6.5 Quakers and their family relationship networks
- 6.6 Immediate relationships
- 6.7 Close relationships
- 6.8 Distant relationships
- 6.9 Significant personal networks - Thomas Hodgkin MD
- 6.10 Significant personal networks - John Hodgkin
- 6.11 Significant personal networks - Edward Backhouse
- 6.12 Significant personal networks - William Fowler
- 6.13 List out all Quaker members of the CEDA
- 6.14 Pie chart Quaker CEDA memberships
- 6.15 Quaker CEDA membership networks
- 6.16 Quaker joiners of the CEDA by years
- 6.17 Quaker leavers of the CEDA by years
- 6.18 Quaker members of the QCA
- 6.19 Quaker joiners of the QCA in years
- 6.20 Quaker leavers of the QCA in years
- 6.21 Show the Quaker members of the APS
- 6.22 Quaker members of the APS - distant relationships
- 6.23 Quaker members of the APS - close relationships
- 6.24 Quaker members of the APS - immediate relationships
- 6.25 Show quaker members of the APS
- 6.26 Pie chart Quaker joiners of the APS
- 6.27 Pie chart Quaker leavers of the APS
- 6.28 Show Quaker members of the ESL

- 6.29 Show Quaker joiners of the ESL
- 6.30 Show Quaker leavers of the ESL
- 6.31 Show Quaker members of the ASL
- 6.32 Show quaker joiners of the ASL
- 6.33 Show quaker leavers of the ASL
- 6.34 Show Quaker members of the AI
- 6.35 Show Quaker joiners of the AI
- 6.36 Show Quaker leavers of the AI
- 6.37 Quaker CEDA members of Case Study Three Hodgkin's network
- 6.38 Quaker CEDA members not in Case Study Three
- P7 Chapter 7 Project Seven presentation
- 7.1 An Evidence Based Prosopographical study of 3000 activists 1830-1870 and the 600 Quakers amongst them
- 7.2 This is a code cell
- 7.3 These are the resources in my container for this exercise
- 7.4 This is the structure of the SQLite database (ERD)
- 7.5 Relationships (other than CEDA) present in the data
- 7.6 All persons are members of at least one CEDA
- 7.7 Quakers and their relationships
- 7.8 Working with a variety of datatables
- 7.9 Introduction to bigraph analysis
- 7.10 Locations
- 7.11 Occupations
- 7.12 Societies
- 7.13 Cubs
- 7.14 Most popular bipartite networks combined
- 7.15 Iterative Section 1 - (This is an iterative workbook)
- 7.16 Listing out the data
- 7.17 Iterative Section 2 - prepare the data for rendering as a graph in Gephi
- 7.18 Stage 2 - Bipartite analysis with 'noise' removed

- 7.19 Simplified graphs can be analysed more easily

Project Seven is a digital study which ran from 2018 to 2024. It is a study of the social networks of 3000 persons who were members of the Centres for the Emergence of the Discipline of Anthropology (CEDA) in Britain 1830-1870, including 600 Quakers amongst them.

The CEDA are:

- The Quaker Committee on the Aborigines, QCA - (1831 – 1846)
- The Aborigines Protection Society, APS - (1837 – 1848)
- The Ethnological Society of London, ESL - (1843 – 1848)
- The Anthropological Society of London, ASL - (1861 – 1869)
- The Anthropological Institute, AI – (1871). A merger of the ESL and ASL

The study used the Human Data Digital Toolkit (HDDT) comprising a SQLite database, Visual Studio Code, Gephi Open Graph Visualisation Platform and Jupyter notebooks (JNB) to analyse and visualise Evidence Based Prosopographical information on the members of the CEDA. The study was performed by Kelvin Beer-Jones working as a lone researcher, Project Seven is a PHD research project for University of Birmingham: Evidence Based Prosopography in the Digital Study of Past Human Lives.

Chapter Summaries:

Chapter 1: Introduces the project and the HDDT

Chapter 2: Introduces the database tables

Chapter 3: Describes the data segmentation for JNB and Gephi analysis

Chapter 4: An analysis of the CEDA memberships

Chapters 5a and 5b: An analysis of the index to Protecting the Empire's Humanity (Laidlaw 2024), the Wellcome Institute London, archive collection The Hodgkin Family Papers (Section D) and the CEDA members

Chapter 6: An analysis of the family relationships between the Quaker members of the CEDA

Chapter 7: A presentation made to the Quaker Studies Research Association 2023.

P7 Chapter 1 – The Human Data Digital Toolkit (HDDT)

File name: jnb_hddt_intro

1.1 The data

My research has revealed the extensive social connectivity between the roughly 3000 members of a 'Quaker Led Group' (QLG). The QLG's activities are spread across five organisations in Britain active between 1830 and 1870, which the Quaker members helped to set up and staff. I call these five organisations, the 'Centres for the Emergence of Discipline of Anthropology in Britain' (CEDA). Amongst the 3000 members of the CEDA are approximately 600 Quakers, 20% of the total membership.

The CEDA comprises:

- The Quaker Committee on the Aborigines, QCA - (1831 – 1846)
- The Aborigines Protection Society, APS - (1837 – 1848)
- The Ethnological Society of London, ESL - (1843 – 1848)
- The Anthropological Society of London, ASL - (1861 – 1869)
- The Anthropological Institute, AI – (1869). A merger of the ESL and the ASL.

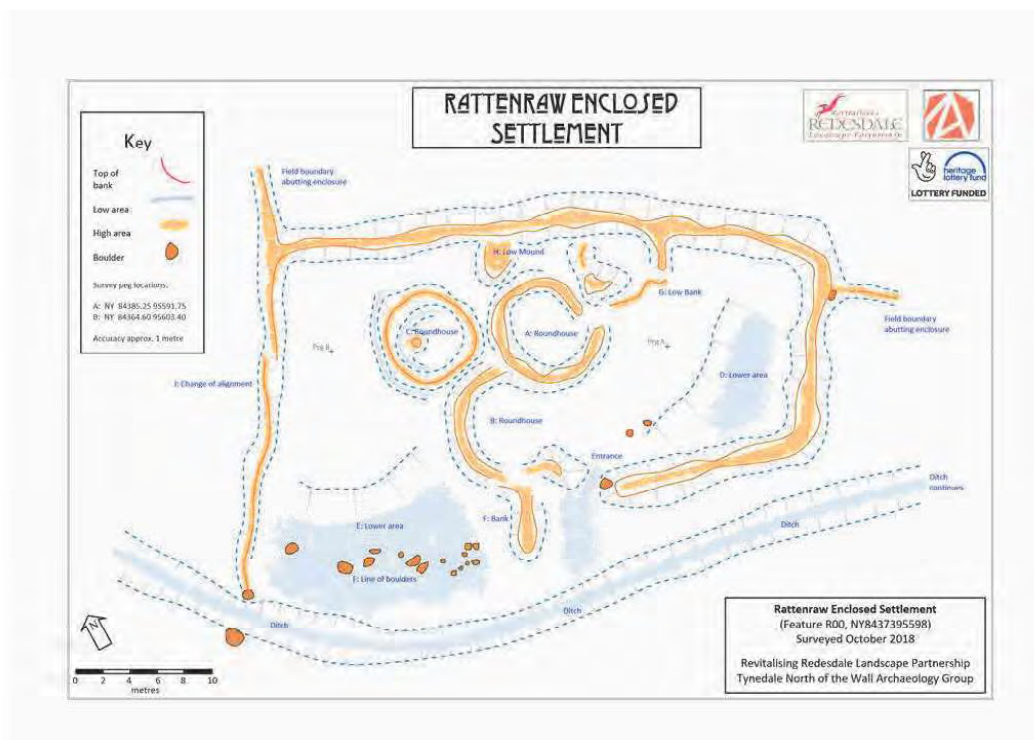
1.2 The HDDT model:

I have designed and built a suite of open-source relational database technologies and digital analytic tools called the Human Data Digital Toolkit (HDDT) to visualise and scrutinise the entire CEDA community over the 40 years of their collective action, from their beginnings in 1830 up to the formation of the Royal Anthropological Institute in 1871. I have identified the Quakers amongst them so that the community can be explored at both whole group and Quaker sub-group levels. Using an SQLite database, Gephi, Visual Studio Code and Jupyter Notebooks I am able to model the prosopographical relationships between the individual members of the CEDA both statically and dynamically (through time). I can analyse bipartite relationships for common attributes: kinship (Quakers only), education, occupations, locations and organisations (whole group). The HDDT allows me to collect, clean, manage and present the data. The model and data have been designed to answer three questions:

HDDT requirements

1. Open source
2. Popular platforms
3. Seamless integration
4. Relational network visualisation capability
5. Semantic Web compatible
6. Extensible technologies
7. Extensible data
8. Reproducible
9. Able to handle the data (quantity and quality)
10. User interventions
11. Supported by University of Birmingham with online learning aids

HTTD vision - to survey data like archaeologists survey sites



To create containers for the data collected in csv sheets from several sources and in different arrangements. To facilitate data combining and cleaning. To establish an auditable data pipeline. To hold the data in an sql database. The model data can then be analysed and visualised by using Gephi which is a compatible open source software package. The HDDT is designed to enable human prosopographical ordered data to be

surveyed, much as an archaeologist might use a variety of technologies to survey a territorial site of historical interest.

The HDDT is a new approach to the Digital Humanities. (1) in its exclusive concentration on Evidence Based Prosopographical data using open source data science techniques readily available to the lone researcher and capable of execution on an average desktop computer. The objective is to facilitate the study large sets of prosopographical data (of variable quality and quantity) either as individuals or groups of persons and embedded sub-groups of persons.

Exponential population growth in the nineteenth century combined with a cultural, disciplined and extensive interest in collecting, cataloguing and preserving historical artefacts and manuscripts, offers to the researcher an opportunity to study individuals, communities and whole sections of societies en masse.

The archives of the nineteenth century are often very large, too big to be surveyed only by the naked eye. (and are immense when viewed collectively). From a prosopographical point of view they offer a rich source of prosopographical data, often organised as metadata (catalogued and frequently cross-referenced, if poorly referenced to primary sources).

The HDDT facilitates the study of prosopographical data in itself, it is purposely free of narratology and interpretation. It, and other tools like it are essential to all researchers wishing to understand the peoples of the past and their relationships.

1.3 Project Seven Questions

Question 1:

Can the model reveal the networks between the members in the five organisations that comprise the CEDA? This question is important because it resolves a wider and current uncertainty over the origins of the discipline of anthropology in Britain and the extent of Quaker involvement.

Question 2

Can the model examine Quaker-to-Quaker relationships and how these relationships supported the Quaker members of the CEDA during the forty years of its life?

Question 3

Can the model reveal the key networking role played by the Quaker Thomas Hodgkin MD (1798–1866) from the beginnings of the CEDA in 1830 up to his death in 1866?

1.4 Contributing datasets

The HDDT integrates ordered datasets from a variety of sources to create one SQLite HTTD database

Historians can create a bespoke database taking data from multiple sources using the HDDT. This project takes data from:

Source	Records
Royal Anthropological Insitiute (RAI)	2260
Quaker Family History Society (QFHS)	593
Independent research at RAI	1171
Independent research at Friends House Quaker Archive, London	30
total records	3095

Table	Col1	Col2	Col3	Col4	Col5
Row1	A	B	C	D	E
Row2	F	G	H	I	J
Row3	K	L	M	N	O
Row4	P	Q	R	S	T
Row5	U	V	W	X	Y

1.5 Dataset variability

Component datasets can be 'Complete', 'Incomplete' or 'Irregular' as long as the contributing datasets consist of records where at least one column is shared. In this Project person name was common to all datasets.

A 'complete' dataset

Would be one like this, where all of the data is contained within a perfect rectangular block of cells ('containers') and every container contains only one data item and every data item can be located by the coordinates 'Row n, Column n'

Table	Col1	Col2	Col3	Col4	Col5
Row1		B	C	D	E
Row2	F		H	I	J
Row3	K	L		N	O
Row4	P	Q	R		T
Row5	U	V	W	X	

An 'incomplete' dataset

When historical data is used often some data is missing (it can be permanently lost). The HDDT is able to accept 'Incomplete' datasets. The HDDT does not lose functionality because of the incomplete nature of much historical data.

1	A		1.0		
2	B		2.1	Cat	Q
3	C		3.2	Dog	W
4	D	fff	4.3	Fish	E
5	E	ggg			R
6	F	hhh			T
7	G	iii			Y

An 'irregular' dataset

The HDDT has been designed to accept Irregular datasets. The surviving evidence of past lives is not only often Incomplete, but also frequently Irregular. An additional complication arises when data from multiple sources must be combined into a single dataset. Here it is likely that multiple donor datasets will have different dimensions. This arises because either the data in itself is intrinsically different or because different data collectors have compiled data at different times and to different standards or simply prefer different collecting methods.

For the HDDT a qualifying contributing dataset is a data set of any dimensions, complete, incomplete or irregular. The only requirement is that all contributing datasets must consist of prosopographical data and with the key field as PERSON-NAME.

1.6 Data pipeline

In the HDDT data transits through a data pipeline with each component of the pipeline managed by a different technology. Great care was taken to ensure the technological array comprised of open source technologies, fully Integratable (losslessly) and widely supported and in wide use in the academic community.

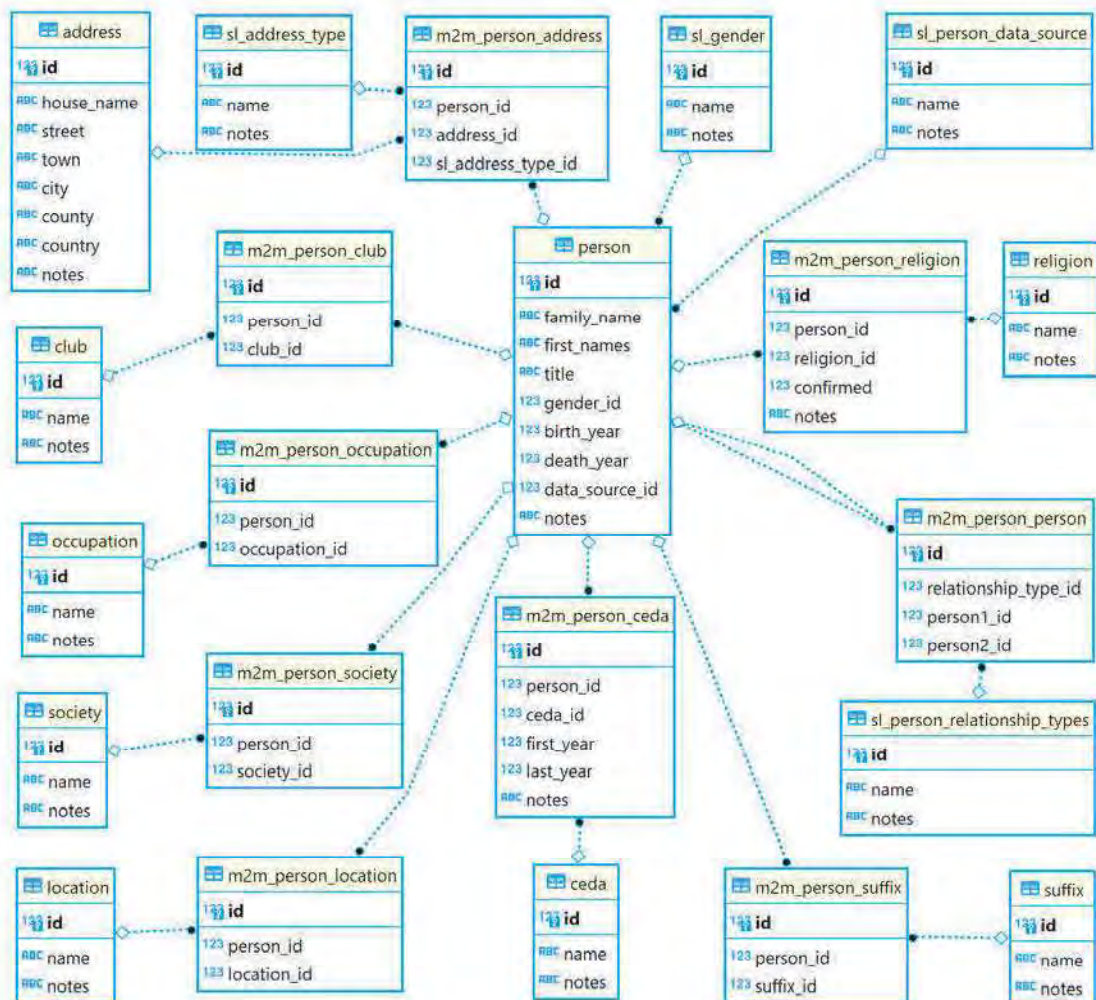


1.7 Model output - Data visualisation criteria

In order to answer the three questions set the selected visualisation technology needed to be able to produce an affiliation network for social network analysis (a bipartite graph). Additionally it needed to be able to present data statically, where all 40 years data are compressed into one visualisation. This would enable the full extent of the network examined to be seen as a single image. Additionally, because the network evolves and changes over time, it was desirable that the technology handle dynamic data, where the network can be visualised progressively one year at a time from 1830 to 1870. Gephi met the project criteria.

1.8 Integrated HTTD SQLite database - 3095 persons

Entity Relationship Diagram (ERD)



At the heart of the HDDT (and the SQLite database) is the 'person table', this holds the data (attributes) unique to each person. All contributing datasets have one or more columns containing person names. Further columns often contain attribute data. Some attribute data may be recorded in several donor datasets. Conflicts between dataset Person Name's are resolved by first choosing to accept the 'RAI dataset' as the 'Authority Index'. Matching names across donor datasets was done by hand in the CSV sheets. This is because (1) person name records from datasets other than the deemed (RAI) dataset were small in number, (2) because name matching requires judgement (names often repeat in families and their might be little attributive data in common between the other donor data set and the RAI dataset and (3) because human matching is common practise in genealogy systems, even if algorithms are used to find possible matches. The finding of possible matches in this project was performed most efficiently by the human eye. (A move to the digital should not needlessly replace the human).

Tables of data items shared amongst persons (such as 'occupation', 'location', 'societies', 'clubs') are linked to the person table by m2m tables. There are also person_person tables to capture family relationships.

1.9 HDDT design architecture

The following packages are required to make the HDDT:

package	Use
GitHub	for version control and sharing
SQL	the database
VSC	building the database
VSC	version control interface to Git
DBeaver	data cleaning, data management and analysis
Jupyter Notebook	data analysis
Gephi	data visualisation

They were chosen because they are universal, popular, open-source and suitable for handling historical ordered data.

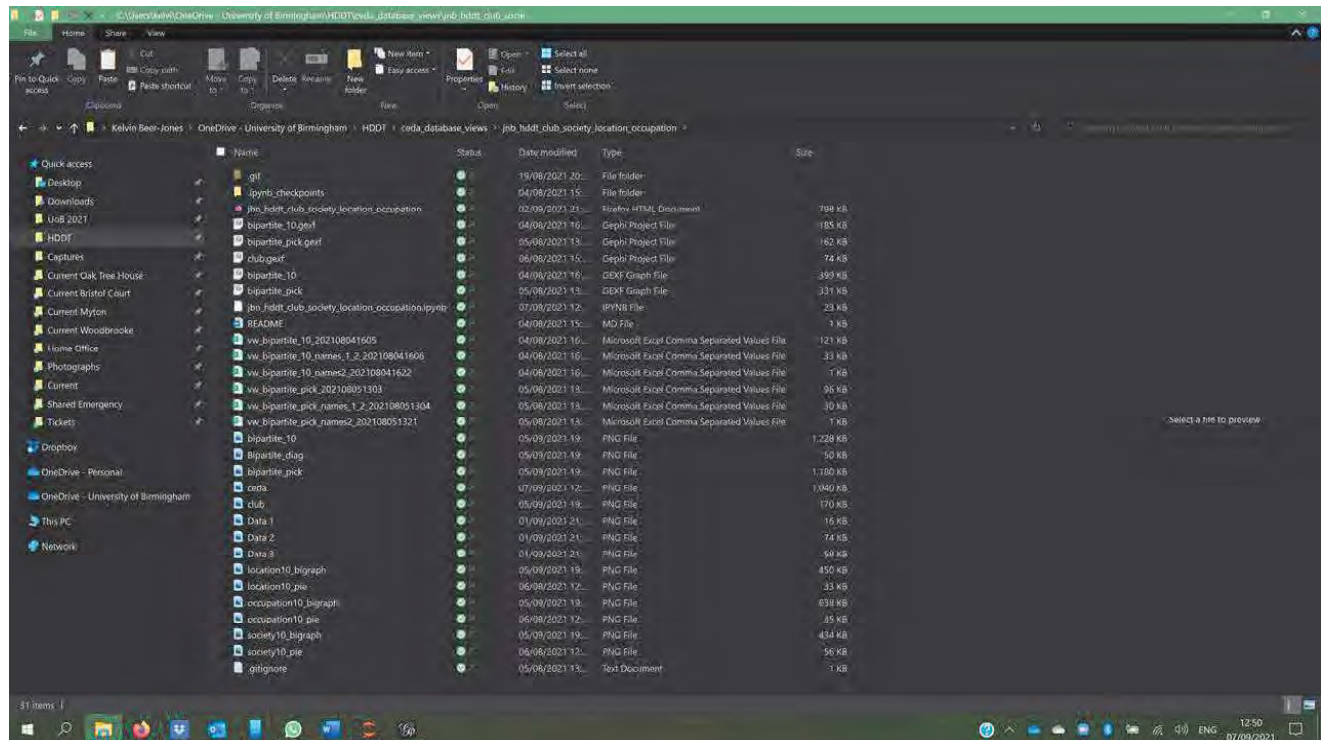
1.10 Project containers must contain all required resources

Data management needs careful consideration and design. The HDDT uses the concept of project containers where every container is set up and initialised as a GitHub repo. Then a Jupyter Notebook is created in the same container. All resources needed for a project are then copied from master containers (such as the template CSV files and dataframes in the ceda/database/views container).

Gexf graph files and gexf project files also are set up and saved in each container.

Relative links can then be used and their integrity preserved.

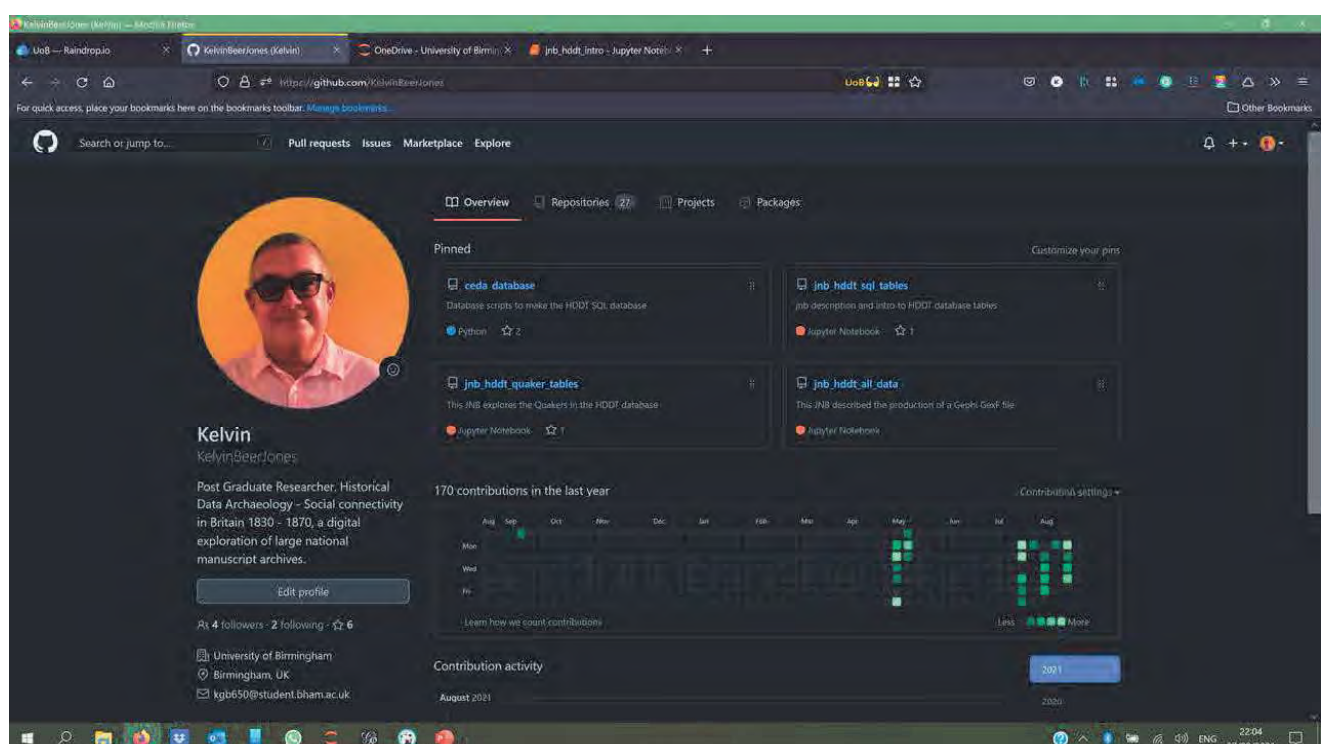
GitHub can also be used for version control providing an audit trail of changes, additions and deletions to the HDDT container system.



1.11 GitHub

The entire HDDT project, its description, structure, organisation and resources are contained in one GitHub account:

[KelvinBeerJones](https://github.com/KelvinBeerJones)



1.12 SQLite recommendations

SQLite is a C-language library that implements a small, fast, self-contained, high-reliability, full-featured, SQL database engine. SQLite is the most used database engine in the world. SQLite is built into all mobile phones and most computers and comes bundled inside countless other applications that people use every day

(<https://www.sqlite.org/index.html>).

SQLite is a C-language library that implements a small, fast, self-contained, high-reliability, full-featured, SQL database engine. SQLite is the most used database engine in the world. SQLite is built into all mobile phones and most computers and comes bundled inside countless other applications that people use every day. The SQLite file format is stable, cross-platform, and backwards compatible and the developers pledge to keep it that way through the year 2050. SQLite database files are commonly used as containers to transfer rich content between systems and as a long-term archival format for data. There are over 1 trillion SQLite databases in active use. SQLite source code is in the public-domain and is free to everyone to use for any purpose.

<https://www.sqlite.org/about.html>

In the two decades following its initial release, SQLite has become the most widely deployed database engine in existence. Today, SQLite is found in nearly every smartphone, computer, web browser, television, and automobile. Several factors are likely responsible for its ubiquity, including its in-process design, standalone codebase, extensive test suite, and cross-platform file format. While it supports complex analytical queries, SQLite is primarily designed for fast online transaction processing (OLTP), employing row-oriented execution and a B-tree storage format. However, fuelled by the rise of edge computing and data science, there is a growing need for efficient in-process online analytical processing (OLAP). DuckDB, a database engine nicknamed “the SQLite for analytics”, has recently emerged to meet this demand. While DuckDB has shown strong performance on OLAP benchmarks, it is unclear how SQLite compares. Furthermore, we are aware of no work that attempts to identify root causes for SQLite’s performance behaviour on OLAP workloads. In this paper, we discuss SQLite in the context of this changing workload landscape. We describe how SQLite evolved from its humble beginnings to the full-featured database engine it is today. We evaluate the performance of modern SQLite on three benchmarks, each representing a different flavour of in-process data management, including transactional, analytical, and blob processing. We delve into analytical data processing on SQLite, identifying key bottlenecks and weighing potential solutions. As a result of our optimizations, SQLite is now up to 4.2X faster on SSB. Finally, we discuss the future of SQLite, envisioning how it will evolve to meet new demands and challenges. ‘Sqlite: past, present, and future’ Gaffney, Kevin P,


1.13 VSC recommendations

Visual Studio Code is a lightweight but powerful source code editor which runs on your desktop and is available for Windows, macOS and Linux. It comes with built-in support for JavaScript, TypeScript and Node.js and has a rich ecosystem of extensions for other languages (such as C++, C#, Java, Python, PHP, Go) and runtimes (such as .NET and Unity) (<https://code.visualstudio.com/>).

‘Visual Studio Code, commonly referred to as VS Code, is an integrated development environment developed by Microsoft for Windows, Linux, macOS and web browsers. Features include support for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded version control with Git. Users can change the theme, keyboard shortcuts, preferences, and install extensions that add functionality. Visual Studio Code is proprietary software released under the “Microsoft Software License”, but based on the MIT licensed program named “Visual Studio Code — Open Source” (also known as “Code — OSS”), also created by Microsoft and available through GitHub. In the 2024 Stack Overflow Developer Survey, out of 58,121 responses, 73.6% of respondents reported using Visual Studio Code, more than twice the percentage of respondents who reported using its nearest text editor and/or IDE alternative, Visual Studio. https://en.wikipedia.org/wiki/Visual_Studio_Code

‘Microsoft develops the free code editor Visual Studio Code (VSCode), that is used by more than 11 million users. In the StackOverflow developer survey from 2019, about 50% of the participants stated that they use VSCode, showing how popular this editor has become. The underlying source code is considered free (as in free speech) and referred to as Code - OSS. Microsoft uses Code - OSS as a base, slightly modifies it (e.g. adds a marketplace integration for distributing plugins), and releases it with a proprietary license under the name “Visual Studio Code”. Although this custom license makes VSCode technically speaking non-free and non-open source, other distributions of Code - OSS are free and contain substitutes for the missing features, for example VSCodium. Further analysis of VSCode and Code - OSS regarding aspects of FOSS (free and open source software) development can be found in the extensive preliminary study that was conducted before the project phase.’ ‘Practical Study of Visual Studio Code’ Michael Plainer (Plainer 2021, 2)

example, in a regular spreadsheet, create analytical reports based on records from different data storages, and export information in an appropriate format. For advanced database users, DBeaver suggests a powerful SQL-editor, plenty of administration features, abilities of data and schema migration, monitoring database connection sessions, and a lot more. Out-of-the box DBeaver supports more than 80 databases. Having usability as its main goal, DBeaver offers:

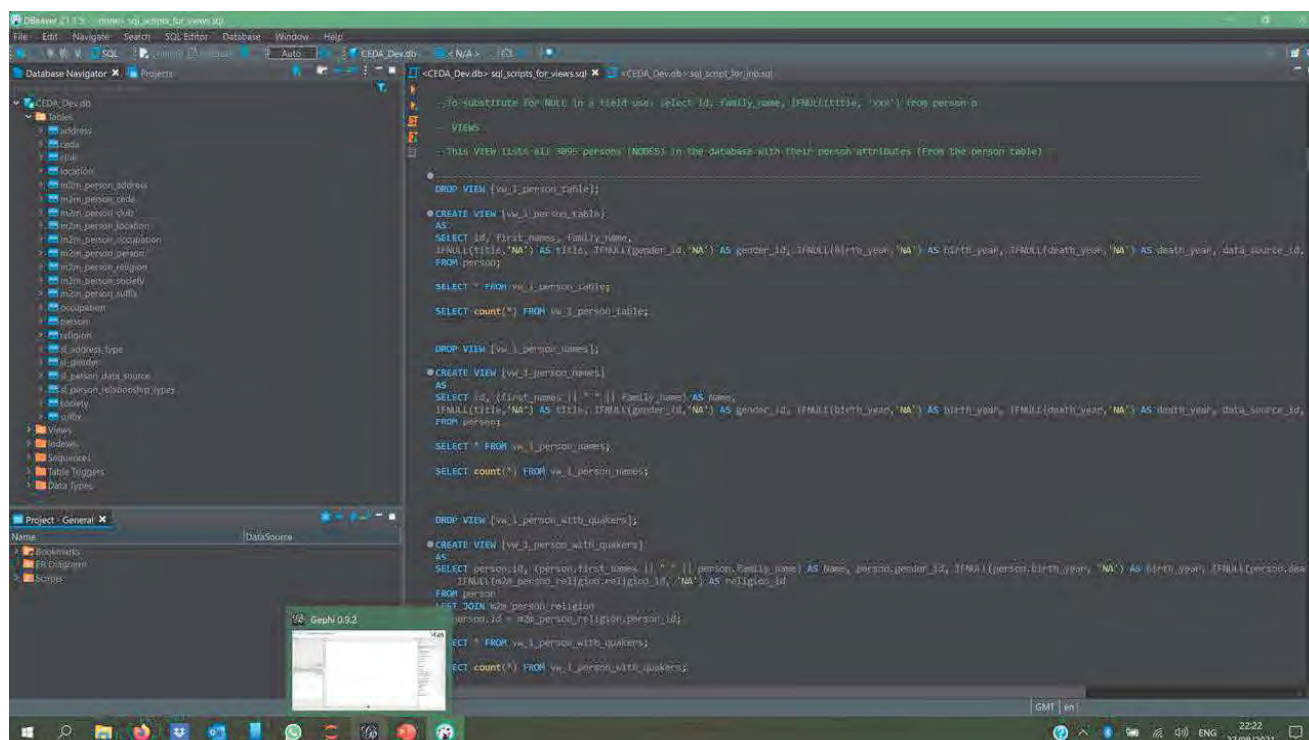
- Carefully designed and implemented User Interface
- Support of Cloud data sources
- Support for Enterprise security standard
- Capability to work with various extensions for integration with Excel, Git, and others.
- Great number of features
- Multiplatform support'  [dbeaver/dbeaver](https://github.com/dbeaver/dbeaver)

DBeaver is a SQL client software application and a database administration tool. For relational databases it uses the JDBC application programming interface (API) to interact with databases via a JDBC driver. For other databases (NoSQL) it uses proprietary database drivers. It provides an editor that supports code completion and syntax highlighting. It provides a plug-in architecture (based on the Eclipse plugins architecture) that allows users to modify much of the application's behavior to provide database-specific functionality or features that are database-independent. It is written in Java and based on the Eclipse platform. The community edition (CE) of DBeaver is a free and open source software that is distributed under the Apache License. A closed-source enterprise edition of DBeaver is distributed under a commercial license.

<https://en.wikipedia.org/wiki/DBeaver>

Universal Database Tool

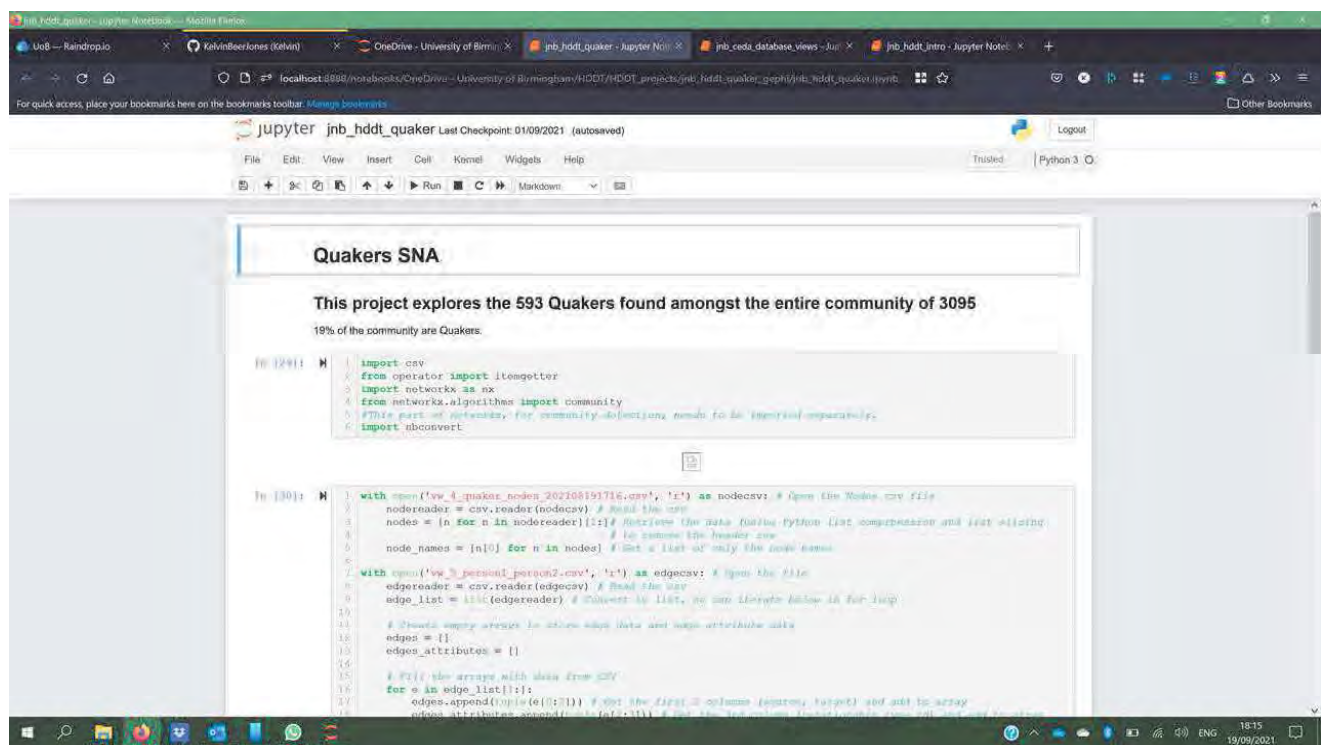
DBeaver is a free multi-platform database tool for developers, database administrators, analysts and all people who need to work with databases. Supports all popular databases: MySQL, PostgreSQL, SQLite, Oracle, DB2, SQL Server, Sybase, MS Access, Teradata, Firebird, Apache Hive, Phoenix, Presto, etc. (<https://dbeaver.io/>)



1.16 Jupyter Notebook recommendations

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning. (<https://jupyter.org/>).

'The Jupyter notebook is an open-source, browser-based tool functioning as a virtual lab notebook to support workflows, code, data, and visualizations detailing the research process. It is machine and human-readable, which facilitates interoperability and scholarly communication. These notebooks can live in online repositories and provide connections to research objects such as datasets, code, methods documents, workflows, and publications that reside elsewhere. Jupyter notebooks are one means to make science more open. Their relevance to the JCDL community lies in their interaction with multiple components of digital library infrastructure such as digital identifiers, persistence mechanisms, version control, datasets, documentation, software, and publications. Our poster examines how Jupyter notebooks embody the FAIR (Findable, Accessible, Interoperable, Reusable) principles for digital objects and assess their utility as viable tools for scholarly communication.' 'Using the Jupyter notebook as a tool for open science: An empirical study' Randles Bernadette M, Pasquetto Irene V, Golshan Milena S, Borgman Christine L. (Randles et al. 2017, 1)



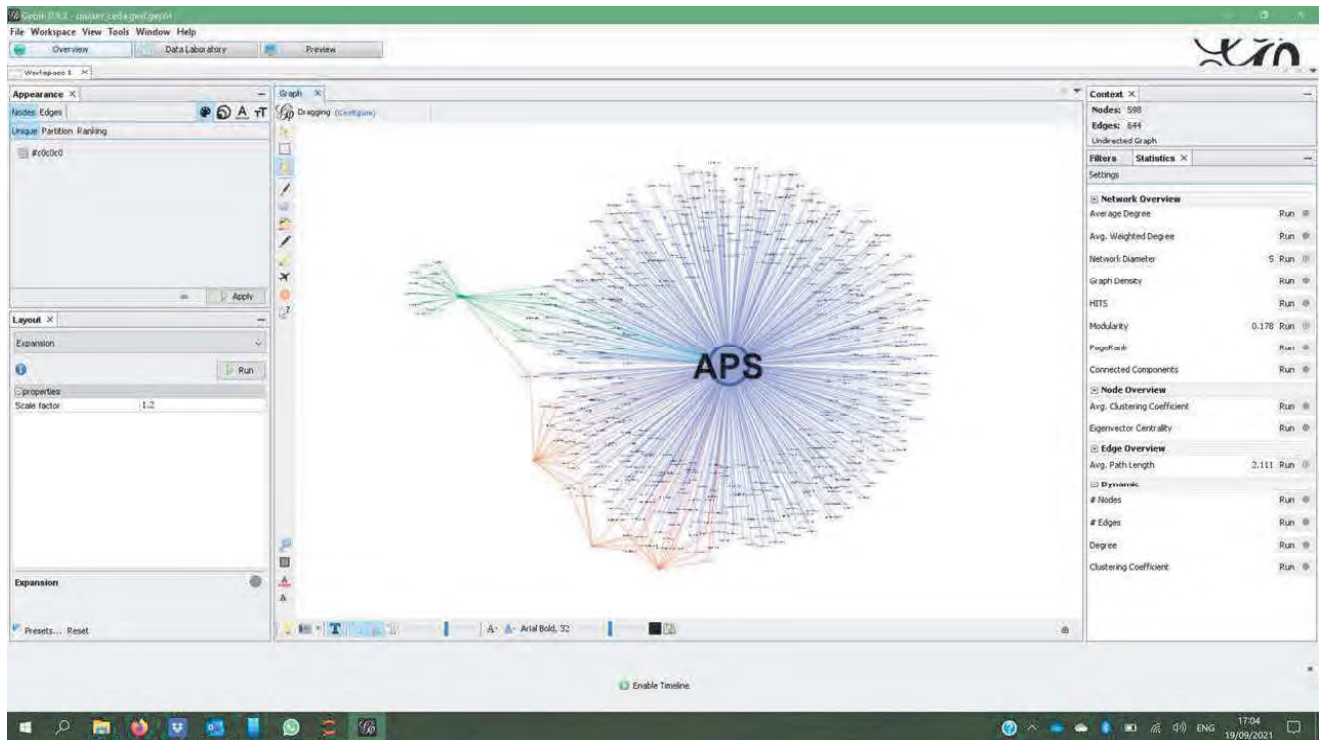
1.17 Gephi recommendations

'Gephi is an open source software for graph and network analysis. It uses a 3D render engine to display large networks in real-time and to speed up the exploration. A flexible and multi-task architecture brings new possibilities to work with complex data sets and produce valuable visual results. We present several key features of Gephi in the context of interactive exploration and interpretation of networks. It provides easy and broad access to network data and allows for spatializing, filtering, navigating, manipulating and clustering. Finally, by presenting dynamic features of Gephi, we highlight key aspects of dynamic network visualization.' Bastian Mathieu, Heymann Sebastien, Jacomy, Mathieu Gephi: an open source software for exploring and manipulating networks. (Bastian, Heymann, and Jacomy 2009, 361)

Gephi is a tool for data analysts and scientists keen to explore and understand graphs. Like Photoshop™ but for graph data, the user interacts with the representation, manipulate the structures, shapes and colors to reveal hidden patterns. The goal is to help data analysts to make hypothesis, intuitively discover patterns, isolate structure singularities or faults during data sourcing. It is a complementary tool to traditional statistics, as visual thinking with interactive interfaces is now recognized to facilitate reasoning. This is a software for Exploratory Data Analysis, a paradigm appeared in the Visual Analytics field of research. <https://gephi.org/features/>

Gephi is a tool for data analysts and scientists keen to explore and understand graphs. Like Photoshop™ but for graph data, the user interacts with the representation,

manipulate the structures, shapes and colors to reveal hidden patterns. The goal is to help data analysts to make hypothesis, intuitively discover patterns, isolate structure singularities or faults during data sourcing. It is a complementary tool to traditional statistics, as visual thinking with interactive interfaces is now recognized to facilitate reasoning. This is a software for Exploratory Data Analysis, a paradigm appeared in the Visual Analytics field of research. (<https://gephi.org/features/>)



1.18 Excel recommendations

<https://www.datacamp.com/tutorial/data-cleaning-in-excel-a-beginners-guide>

https://cartong.pages.gitlab.cartong.org/learning-corner/en/3_nettoyage/3_3_nettoyage_donnees

1.19 References

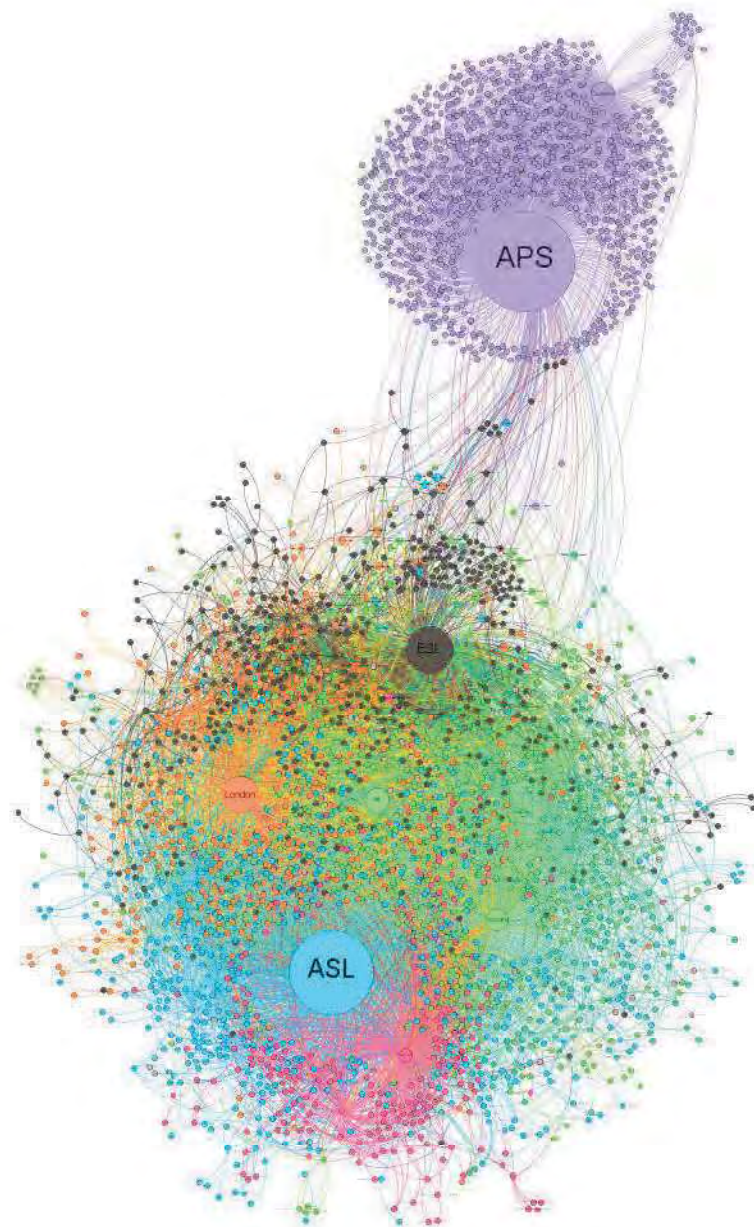
Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. 2009. "Gephi: an open source software for exploring and manipulating networks." Proceedings of the international AAAI conference on web and social media.

Gaffney, Kevin P, Martin Prammer, Larry Brasfield, D Richard Hipp, Dan Kennedy, and Jignesh M Patel. 2022. "Sqlite: past, present, and future." Proceedings of the VLDB Endowment 15 (12).

Plainer, Michael. 2021. "Practical Study of Visual Studio Code."

Randles, Bernadette M, Irene V Pasquetto, Milena S Golshan, and Christine L Borgman. 2017. "Using the Jupyter notebook as a tool for open science: An empirical study." 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL).

1.20 600 Quakers amongst 3000 activists for 40 years 1830-1870



P7 Chapter 2 The CEDA Members

Install necessary resources

```
import pandas as pd

import csv

# First call the sql tables as they appear in the database (Part One)

ceda = pd.read_csv ('ceda_202107151833.csv')
club = pd.read_csv ('club_202107151833.csv')
location = pd.read_csv ('location_202107151833.csv')
person_ceda = pd.read_csv ('m2m_person_ceda_202107151835.csv')
person_club = pd.read_csv ('m2m_person_club_202107151835.csv')
person_location = pd.read_csv ('m2m_person_location_202107151835.csv')
person_occupation = pd.read_csv ('m2m_person_occupation_202107151836.csv')
person_person = pd.read_csv ('m2m_person_person_202107151836.csv')
person_religion = pd.read_csv ('m2m_person_religion_202107151836.csv')
person_society = pd.read_csv ('m2m_person_society_202107151836.csv')
occupation = pd.read_csv ('occupation_202107151834.csv')
person = pd.read_csv ('person_202107151834.csv')
religion = pd.read_csv ('religion_202107151834.csv')
society = pd.read_csv ('society_202107151834.csv')

# next call the sql tables rendered in Gephi format (Part Two)

# call all Names

gephi_all_names = pd.read_csv ('vw_2_all_bipartite_memberships_202107121854.csv')
gephi_names_notceda = pd.read_csv ('vw_2_all_bipartite_memberships_xceda_202107121854.csv')

# Then call all Tuples (Source and Target)

gephi_person_ceda = pd.read_csv ('vw_2_ceda_membership_202107121855.csv')
gephi_person_club = pd.read_csv ('vw_2_club_membership_202107121855.csv')
gephi_person_location = pd.read_csv ('vw_2_location_membership_202107121856.csv')
gephi_person_occupation = pd.read_csv ('vw_2_occupation_membership_202107121856.csv')
gephi_person_person = pd.read_csv ('vw_2_person_person_relationships_202107121856.csv')
gephi_person_religion = pd.read_csv ('vw_2_religion_membership_202107121856.csv')
gephi_person_society = pd.read_csv ('vw_2_society_membership_202107121858.csv')

import matplotlib.pyplot as plt

plt.rcParams.update({'font.size': 18})

plt.rc('figure', figsize=(20, 10))

import numpy as np
```



```

-----
FileNotFoundError                                Traceback (most recent call last)
Cell In[1], line 7
      3 import csv
      5 # First call the sql tables as they appear in the database (Part One)
----> 7 ceda = pd.read_csv ('ceda_202107151833.csv')
      8 club = pd.read_csv ('club_202107151833.csv')
      9 location = pd.read_csv ('location_202107151833.csv')

File /Library/Frameworks/Python.framework/Versions/3.10/lib/python3.10/site-
 1013 kwds_defaults = _refine_defaults_read(
 1014     dialect,
 1015     delimiter,
 1016     (...)
 1022     dtype_backend=dtype_backend,
 1023 )
 1024 kwds.update(kwds_defaults)
-> 1026 return _read(filepath_or_buffer, kwds)

File /Library/Frameworks/Python.framework/Versions/3.10/lib/python3.10/site-
 617 _validate_names(kwds.get("names", None))
 619 # Create the parser.
--> 620 parser = TextFileReader(filepath_or_buffer, **kwds)
 622 if chunksize or iterator:
 623     return parser

File /Library/Frameworks/Python.framework/Versions/3.10/lib/python3.10/site-
 1617     self.options["has_index_names"] = kwds["has_index_names"]
 1619 self.handles: IOHandles | None = None
-> 1620 self._engine = self._make_engine(f, self.engine)

File /Library/Frameworks/Python.framework/Versions/3.10/lib/python3.10/site-
 1878     if "b" not in mode:
 1879         mode += "b"
-> 1880 self.handles = get_handle(
 1881     f,
 1882     mode,
 1883     encoding=self.options.get("encoding", None),
 1884     compression=self.options.get("compression", None),
 1885     memory_map=self.options.get("memory_map", False),
 1886     is_text=is_text,
 1887     errors=self.options.get("encoding_errors", "strict"),
 1888     storage_options=self.options.get("storage_options", None),
 1889 )
 1890 assert self.handles is not None
 1891 f = self.handles.handle

File /Library/Frameworks/Python.framework/Versions/3.10/lib/python3.10/site-
 868 elif isinstance(handle, str):
 869     # Check whether the filename is to be opened in binary mode.
 870     # Binary mode does not support 'encoding' and 'newline'.
 871     if ioargs.encoding and "b" not in ioargs.mode:
 872         # Encoding
--> 873         handle = open(
 874             handle,
 875             ioargs.mode,
 876             encoding=ioargs.encoding,
 877             errors=errors,
 878             newline="",

```



```
879         )
880     else:
881         # Binary mode
882         handle = open(handle, ioargs.mode)
```

```
FileNotFoundError: [Errno 2] No such file or directory: 'ceda_202107151833.c'
```

2.1 Introduction

The Historical Data Digital Toolkit (HDDT) comprises of data extracted from one or more ordered datasets (which can be data sourced from several archives). Multiple donor datasets must share a common data field to enable datasets from multiple donors to be linked together (combined). For this project the common data field is a person's name. Data extracted from the RAI, QFHS and my own research (at RAI and Friends House, London), produced records where each record provided data about a person. All of the persons in this project were members of the Centres for the Emergence of the Discipline of Anthropology in Britain 1830 to 1870 (the CEDA).

Collected data (after cleaning and combining) was then rendered as CSV sheets and these were used to create SQL database tables.

The person columns 'Family Name' and 'First Names' in the database are common to all datatables. Therefore the person table is the 'master' table (all persons recorded have a unique ID, and authority index concerns were resolved by accepting the RAI dataset as the project authority index). Attributable data was then attached to the person table (such as date of birth).

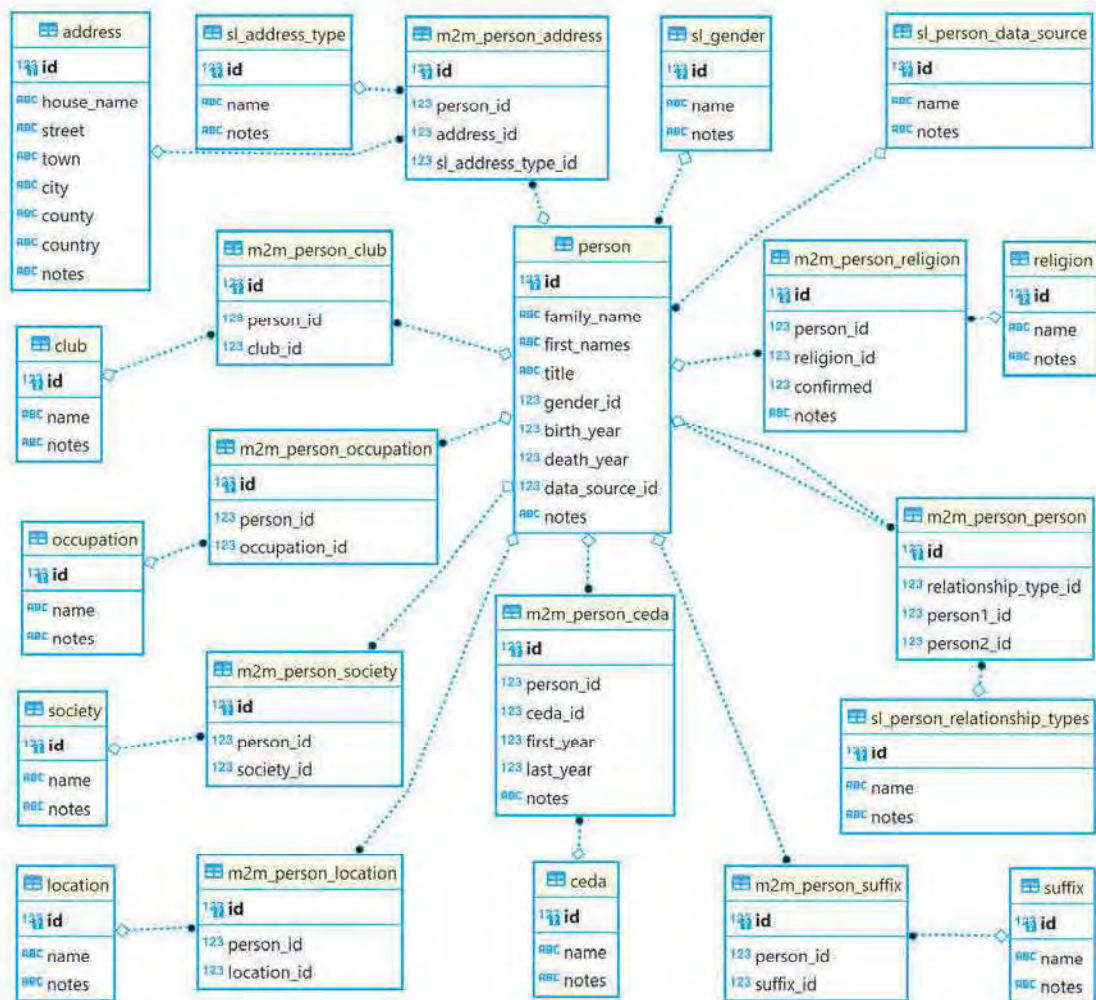
Other related data items are captured in relatable tables and all relatable tables have a shared structure ('id', 'name', 'notes').

m2m tables were then built linking the person data table to related datatables. m2m tables allow for many to many relationships.

Note: The m2m_person_ceda table includes the attributable data 'first_year' and 'last_year'.

The person_person table will generate social network graphs, all other m2m_tables will generate bipartite graphs (bigraphs).

2.2 The Entity Relationship Diagram



2.3 ERD statistics

6 CEDA Societies:

1. QCA Quaker Committee on the Aborigines, 31 members.
2. APS Aborigines Protection Society, 1171 members.
3. ESL Ethnological Society of London, 748 members.
4. ASL Anthropological Society of London, 1334 members.
5. AI Anthropological Institute. 610 members.

514 attribute types:

- ceda (6)

- club (68)
- location (83)
- occupation (93)
- person (3095)
- suffix (155)
- religion (4) Only one group is present, 1 = Quaker.
- society (260)

12097 relationships

- m2m_person_ceda (3894)
- m2m_person_club (323)
- m2m_person_location (2061)
- m2m_person_occupation (1883)
- m2m_person_person (2099)
- m2m_person_religion (593)
- m2m_person_society (1238)
- m2m_person_suffix (1351) *Not used in this project*

2.4 The structure and dimensions of all SQL

tables

Table	Rows	Columns
ceda	6	1
person_ceda	3894	4
club	68	1
person_club	323	2
location	83	1
person_location	2061	2
occupation	93	1
person_occupation	1883	2
person	3095	7
person_person	2099	3
religion	4	1
person_religion	593	3
society	260	1
person_society	1238	2

Note:

Bipartite relationships = 9992 (+2099 person_person relationships = 12091 total relationships)

2.5 The CEDA

code cells

ceda

	id	name	notes
0	1	QCA	NaN
1	2	APS	NaN
2	3	ESL	NaN
3	4	ASL	NaN
4	5	LAS	NaN
5	6	AI	NaN

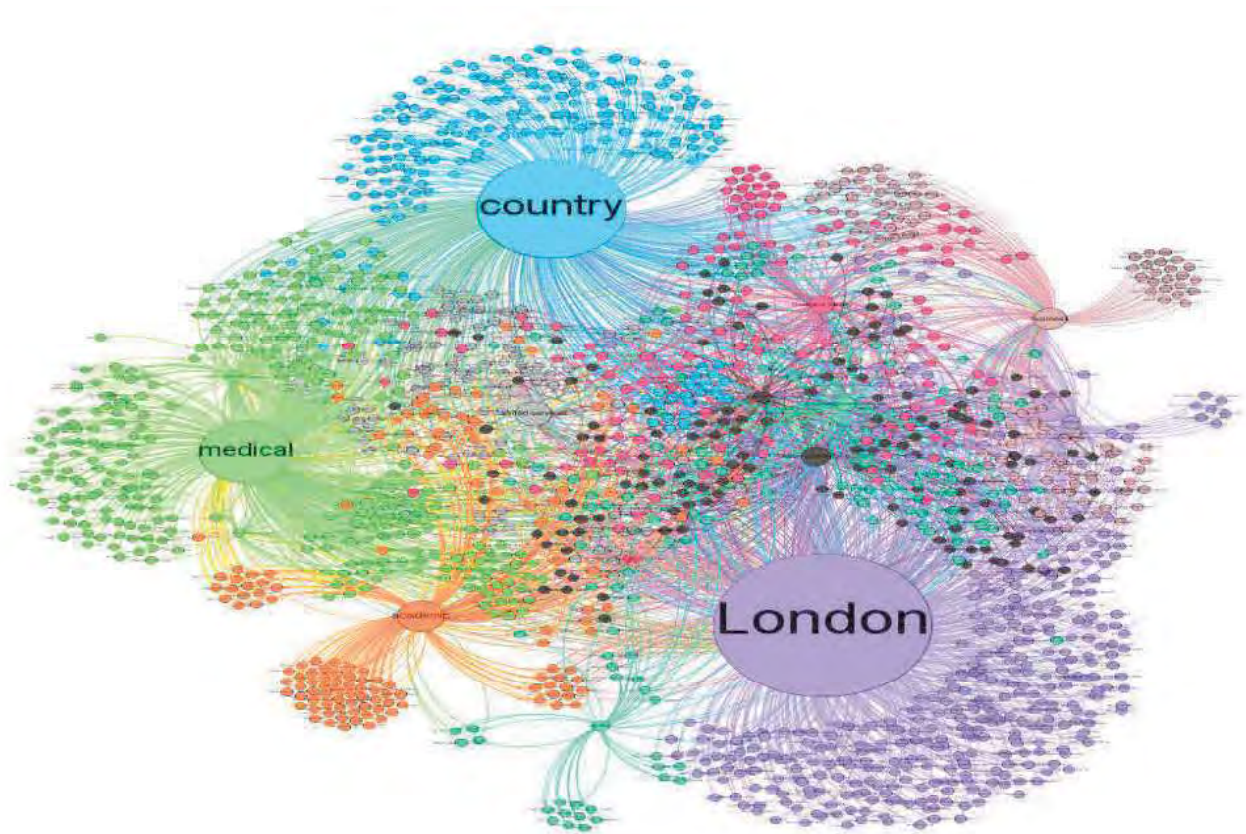
person_ceda

	id	person_id	ceda_id	first_year	last_year	notes
0	1	5	3	1844	1844	NaN
1	2	7	3	1844	1844	NaN
2	3	8	3	1858	1871	NaN
3	4	12	3	1860	1871	NaN
4	5	14	3	1843	1845	NaN
...
3889	4096	3415	2	1839	1850	NaN
3890	4097	3416	2	1861	1862	NaN
3891	4098	3417	2	1853	1856	NaN
3892	4099	3418	2	1840	1840	NaN
3893	4100	3419	2	1840	1867	NaN

3894 rows x 6 columns

2.6 All CEDA members' relationships

visualisation



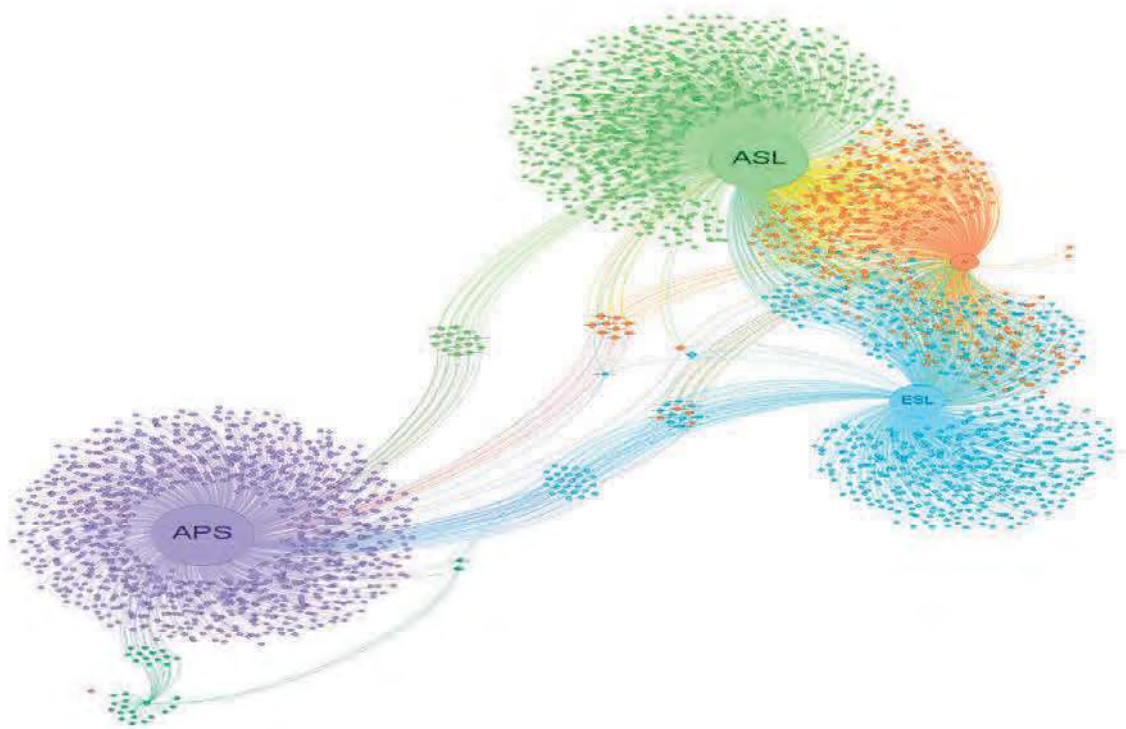
The graph above shows all of the popular bigraph data in the database. Including all data result in a 'hairball' because the network is too dense to be capable of analysis at this (the highest) level.

1850 members of the community are recorded as members of 35 popular entities (Locations, occupations, societies and the Athenaeum Club). These entities make a sphere of popular interest graph where meetings between members concerning the CEDA may have taken place, equally they may also be places where members might meet up only infrequently or informally.

The visual analysis of connectivity between members in single societies and between members of multiple societies indicates the extent that the community is societally connected. The 1850 make up 60% of the entire community.

2.7 All persons are members of at least one

CEDA society



2.8 Memberships in each CEDA table

```
person_ceda.groupby('ceda_id')['person_id'].nunique().plot(kind='bar')
plt.title ("CEDA memberships")
plt.xlabel ("CEDA")
plt.ylabel ("Number of persons")
plt.show()
```

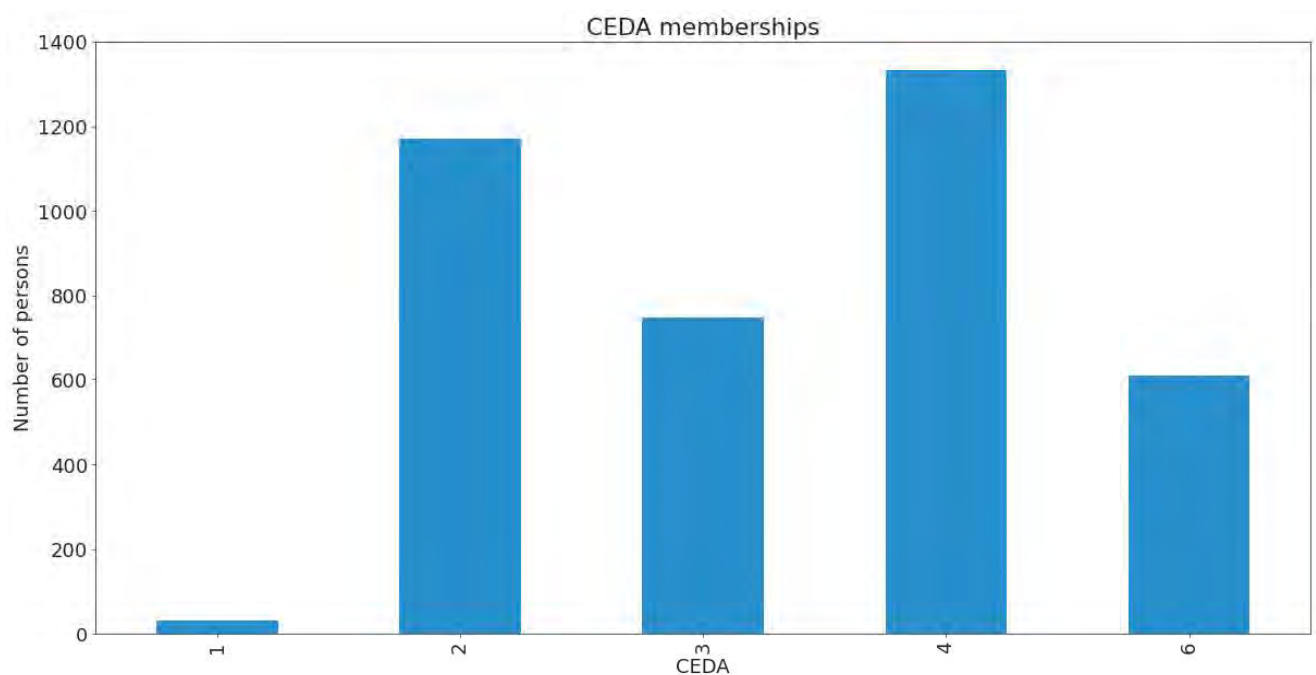


Chart (above) Memberships by CEDA society. - 1 QCA, - 2 APS, - 3 ESL, - 4 ASL, - 6 AI

Note: 5 - LAS data has not been collected by the RAI archivists

2.9 CEDA members were also members of 68 clubs

club

	id	name	notes
0	2	Athenaeum Club	NaN
1	3	Marlborough Club	NaN
2	4	Carlton Club	NaN
3	5	Oriental Club	NaN
4	6	National Club	NaN
...
63	65	Royal Albert Yacht Club	NaN
64	66	Berwickshire Naturalists Field Club	NaN
65	67	Indian Club	NaN
66	68	Ad Eundem	NaN
67	69	Arthurs Club	NaN

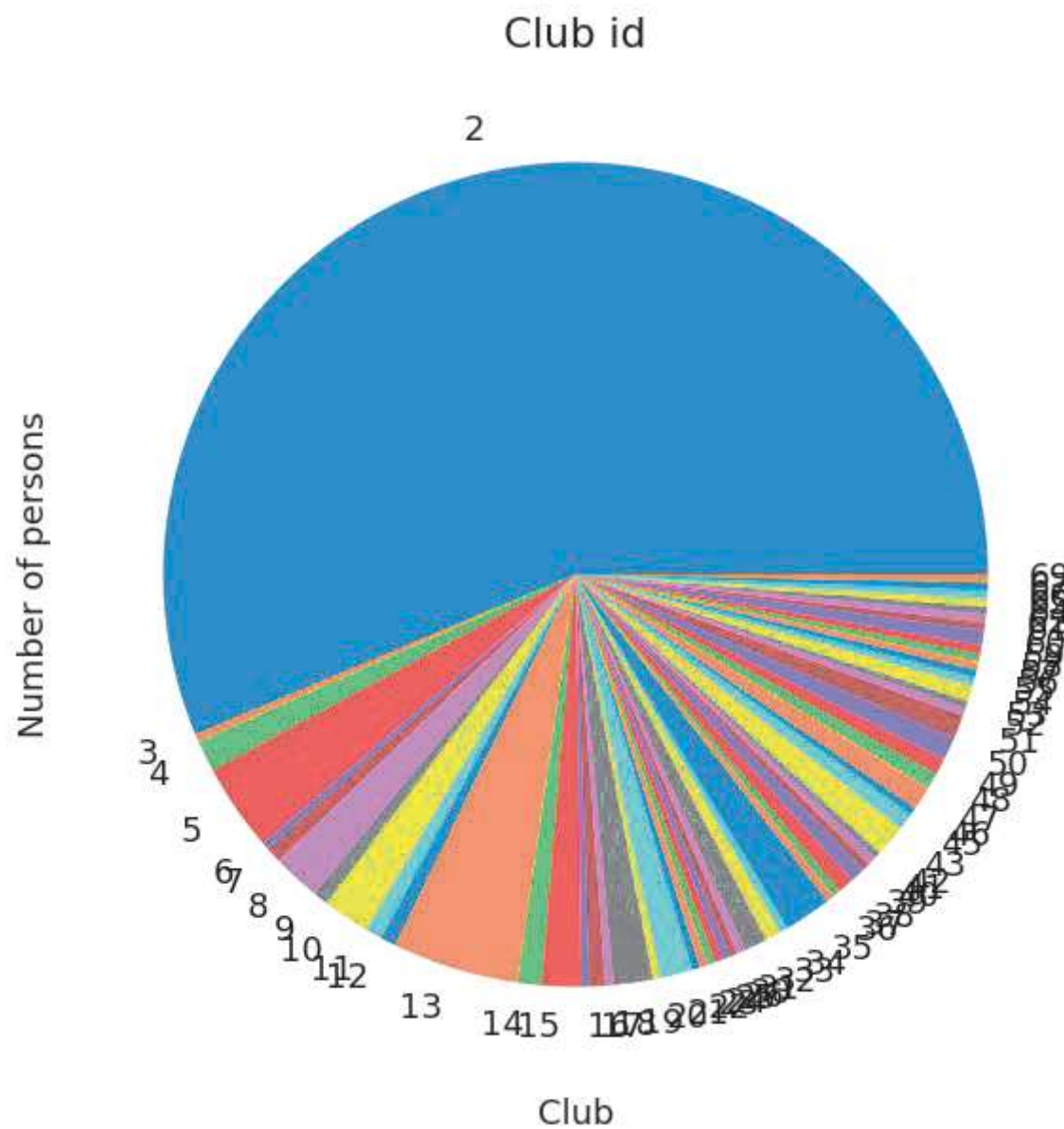
68 rows x 3 columns

person_club

	id	person_id	club_id
0	1	7	2
1	2	22	2
2	3	33	2
3	4	33	3
4	5	33	4
...
318	356	2163	5
319	357	2196	2
320	358	2214	4
321	359	2223	2
322	360	2251	2

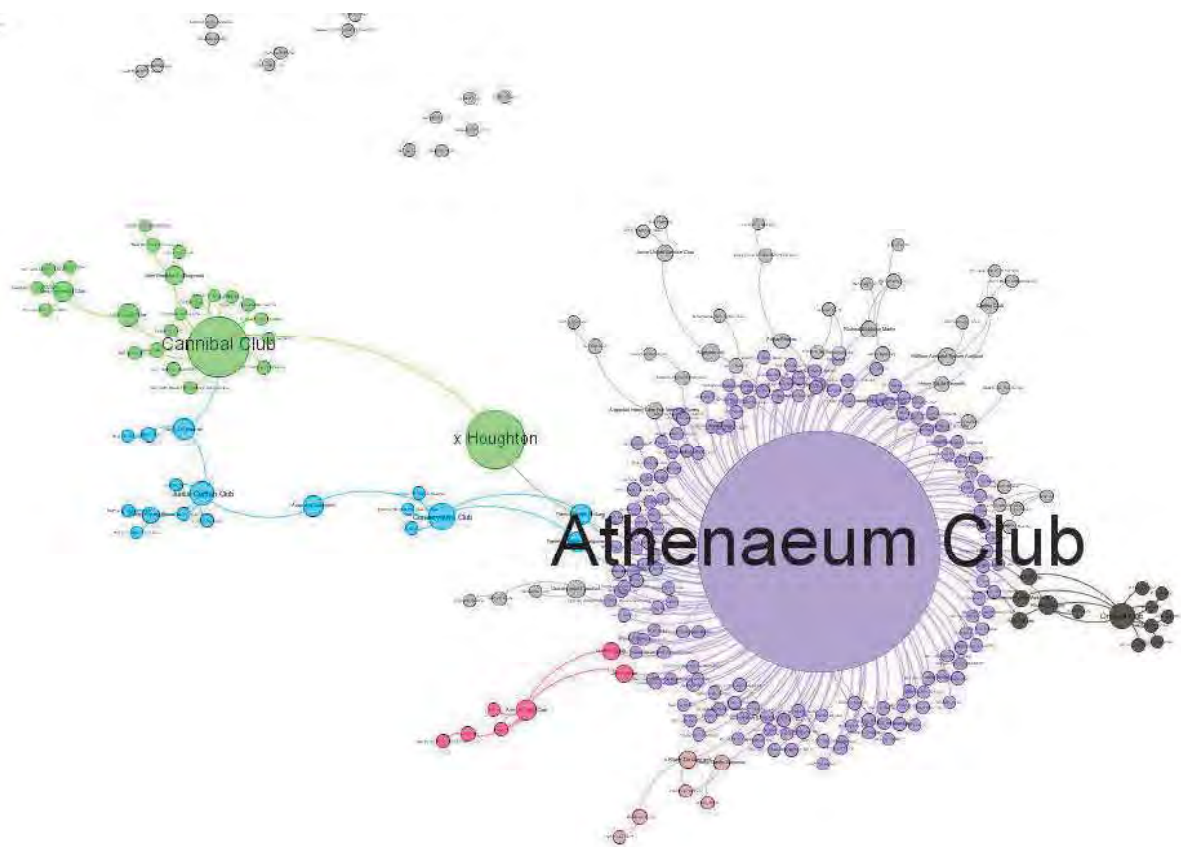
323 rows x 3 columns

```
person_club.groupby('club_id')['person_id'].nunique().plot(kind='pie')
plt.title ("Club id")
plt.xlabel ("Club")
plt.ylabel ("Number of persons")
plt.show()
```



The Athaneum Club membership (2) exceeds that of all other clubs combined.

Note - ignore clubs outside of the top 5?



Clubs will not be analysed in this project but the Athenaeum club can be used as an attribute (because it is a singularity).

2.10 CEDA members are identified with 83 locations

location

	id	name	notes
0	1	London	NaN
1	3	country	NaN
2	4	Africa	NaN
3	5	America	NaN
4	6	Scotland	NaN
...
78	80	Madagascar	NaN
79	81	Ecuador	NaN
80	82	Seychelles	NaN
81	83	Panama	NaN
82	84	Armenia	NaN

83 rows × 3 columns

person_location

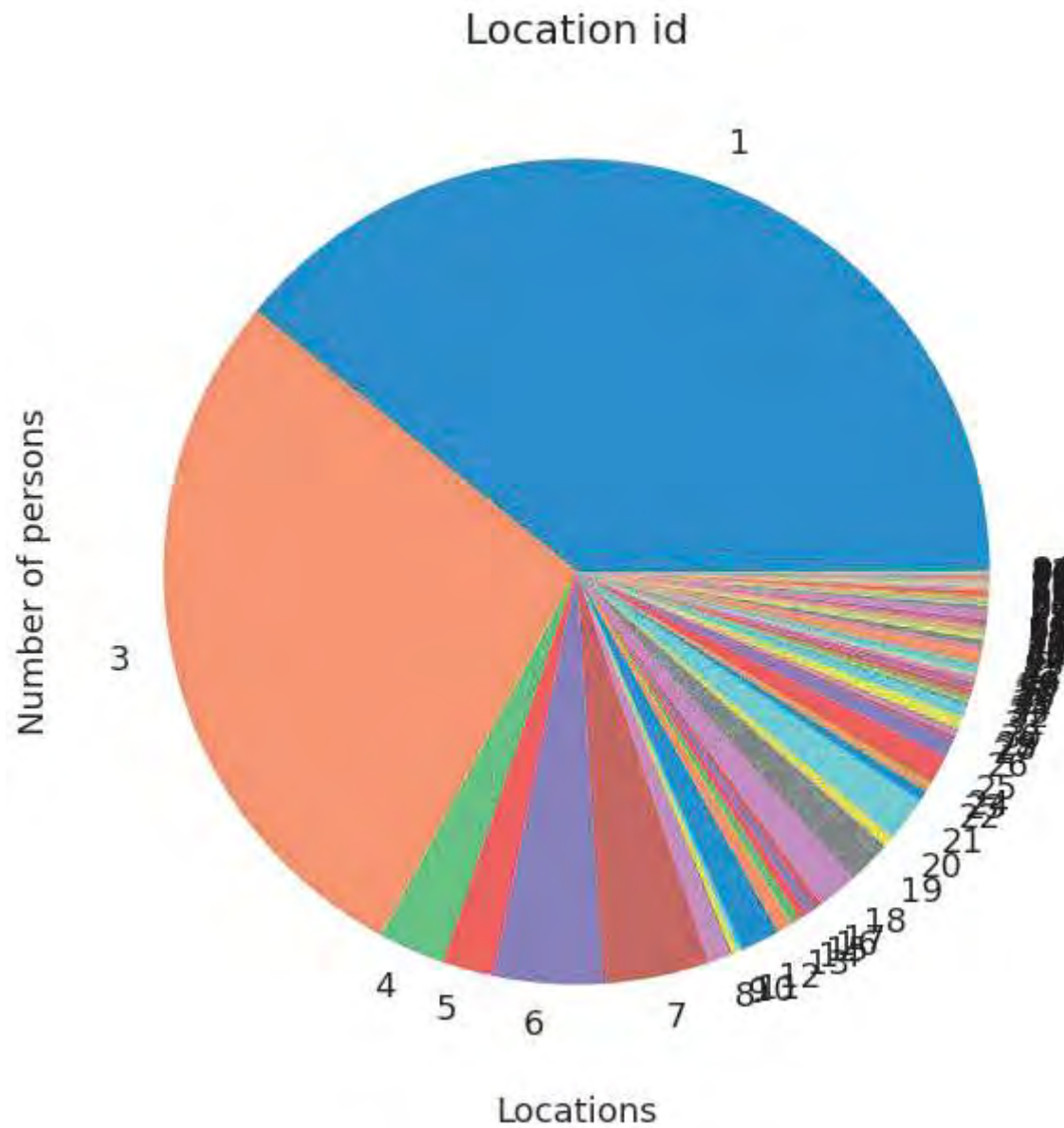
	id	person_id	location_id
0	1	1	1
1	2	3	3
2	3	4	4
3	4	6	1
4	5	8	1
...
2056	2257	2255	1
2057	2259	2258	3
2058	2260	2259	83
2059	2261	2260	84
2060	2262	2260	31

2061 rows × 3 columns

```

person_location.groupby('location_id')['person_id'].nunique().plot(kind='pie')
plt.title ("Location id")
plt.xlabel ("Locations")
plt.ylabel ("Number of persons")
plt.show()

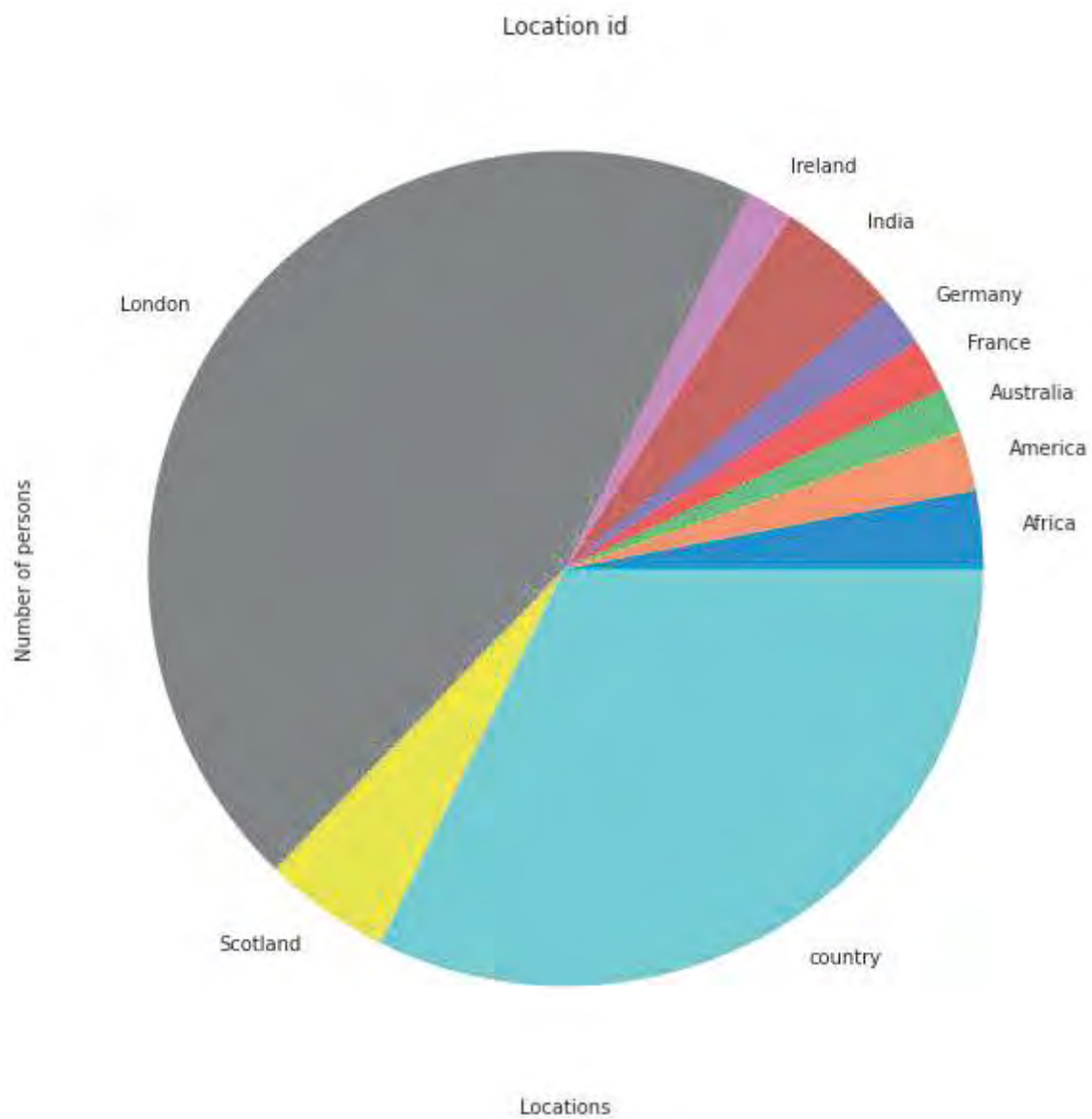
```



Location ID	location	Count
1	London	806
3	country	578
6	Scotland	88
7	India	86
4	Africa	54
5	America	41
19	Germany	37
21	France	36
12	Ireland	32
18	Australia	32
25	Wales	23

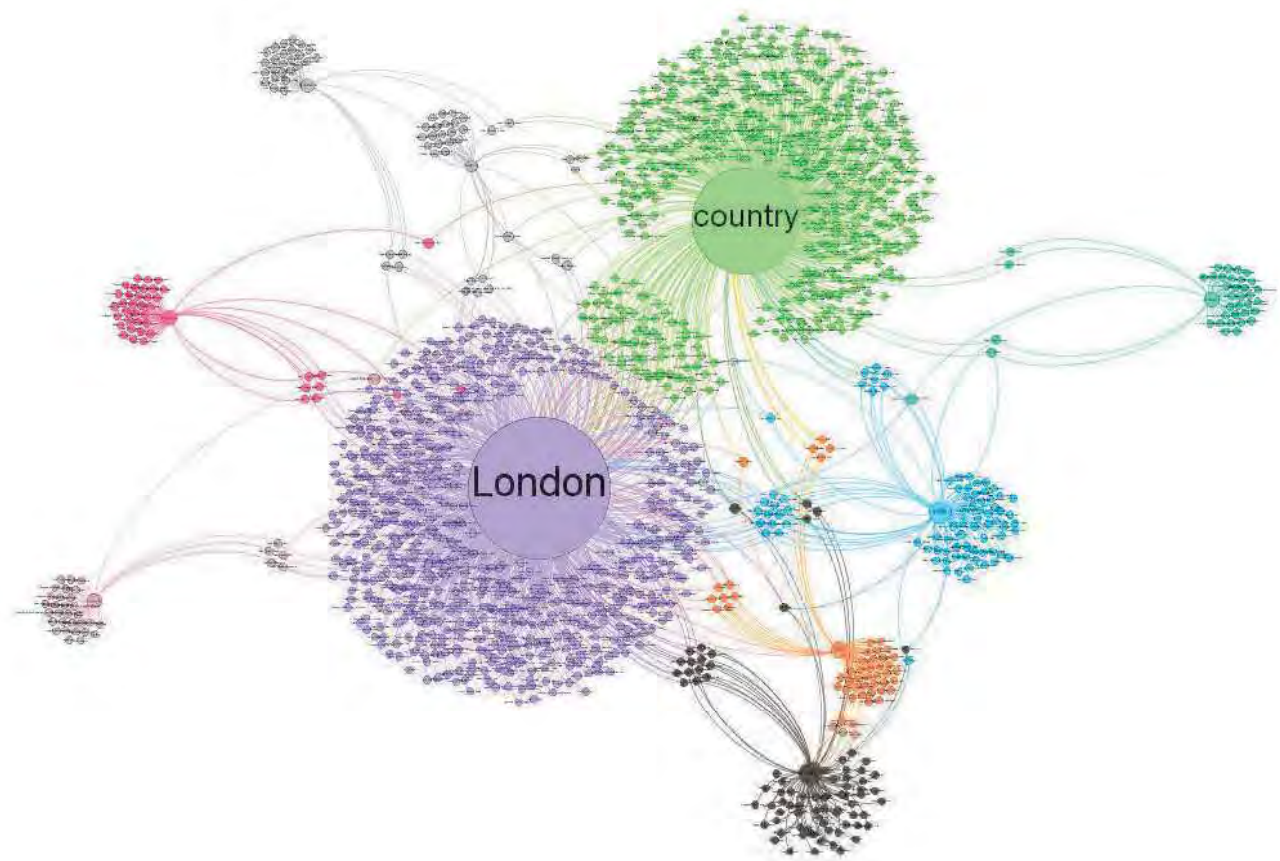
Locations table. Note: London, Country and Scotland equal 1472, more than 50% of all locations

2.11 top 10 locations



2.12 Some persons are associated with

multiple locations



We can see that London and 'country' (sic) are the most populated locations. Because the 'country' location is an aggregate (and not a specific location) we can think of London and 'country' as a twin centre. Within the twin centre we can see the members of both London and 'country' locations and that the members of each are highly networked. We can also see that the London location contains many members who have no association with any other group (including 'country'). London 1830 - 1870, was densely populated and so it is possible that members of the London location had other modes of association. Because the 'country' location is an aggregate we cannot make the same analysis to the same extent, it is possible that many members in (say) Newcastle had no association with other members in (say) Bristol. We can see the large group of members who were members of both London and 'country' locations. It is highly likely that these members served as conduits of communication and group cohesion. It is interesting to note that only 3 members of this London and 'country' group were members of groups outside of the twin centre.

Eight other location each have a membership of around 30 members (we can call these the satellites), all of the satellite groups relate directly to the twin centre with very few members associated with more than one satellite location.

Australia and Ireland have associations with both London and 'country'. The German location is most closely associated with the 'country' group. All of the other locations are strongly associated with the London location.

Germany (far right) is the location least associated with London. Alex Nidda Genthe is the only member from Germany who is also a member of the London location. Friedrich Max Muller, Frederick Augustus Haverick and Gustav Oppert each network with 'country' members. William Wilson Hunter is the only 'country' member who also appears in the Germany location. He and Gustav Oppert also have a location connection with India.

2.13 CEDA members occupations

occupation

	id	name	notes
0	1	literary	NaN
1	3	medical	NaN
2	4	armed services	NaN
3	5	political	NaN
4	6	church	NaN
...
88	90	farmer	NaN
89	91	clockmaker	NaN
90	92	plant collector	NaN
91	93	private means	NaN
92	94	oceanographer	NaN

93 rows x 3 columns

person_occupation

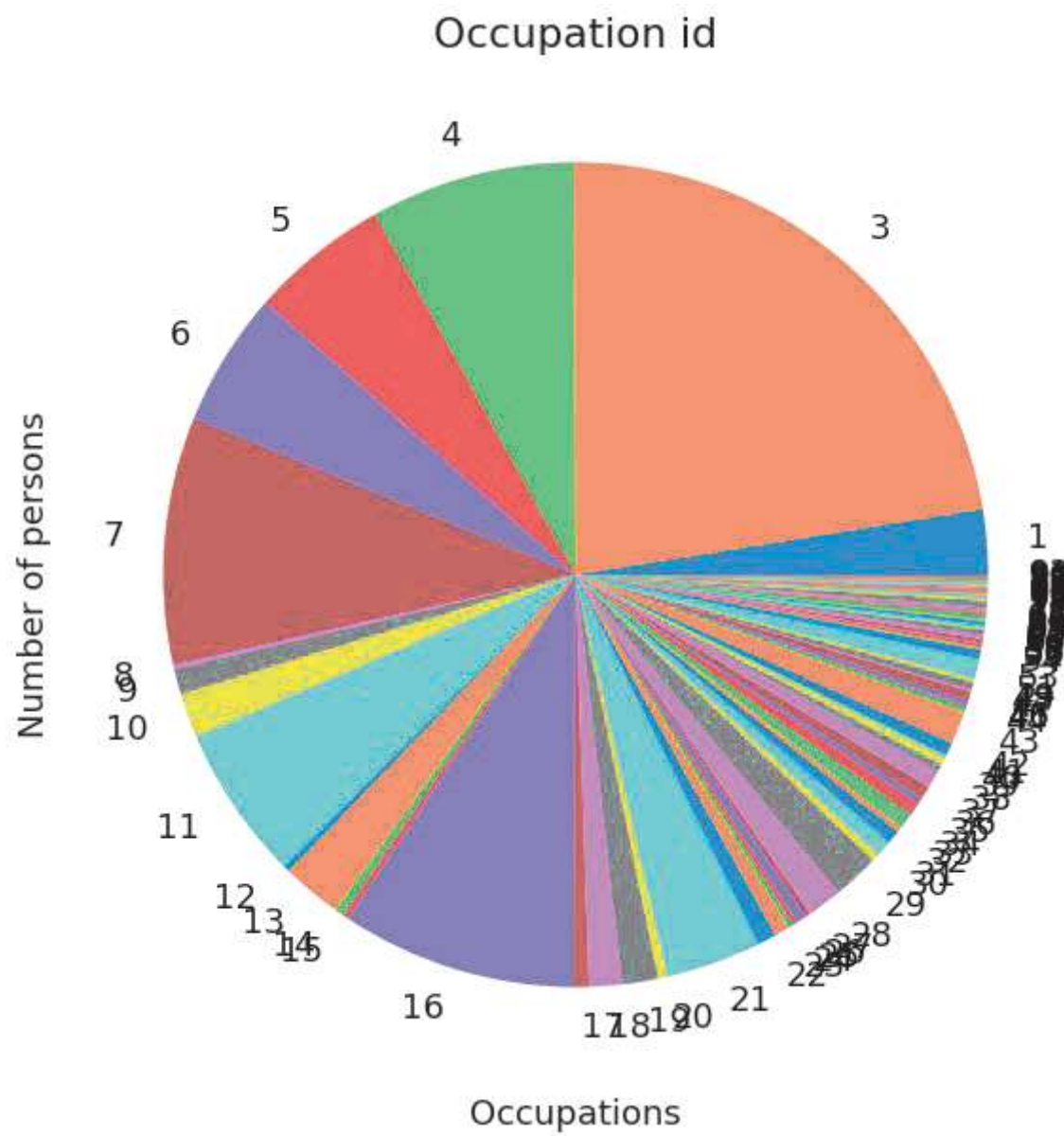
	id	person_id	occupation_id
0	1	1	1
1	2	3	3
2	3	3	4
3	4	5	5
4	5	8	3
...
1878	2122	2252	3
1879	2123	2253	3
1880	2124	2254	3
1881	2125	2255	16
1882	2127	2259	13

1883 rows x 3 columns

```

person_occupation.groupby('occupation_id')['person_id'].nunique().plot(kind=
plt.title ("Occupation id")
plt.xlabel ("Occupations")
plt.ylabel ("Number of persons")
plt.show()

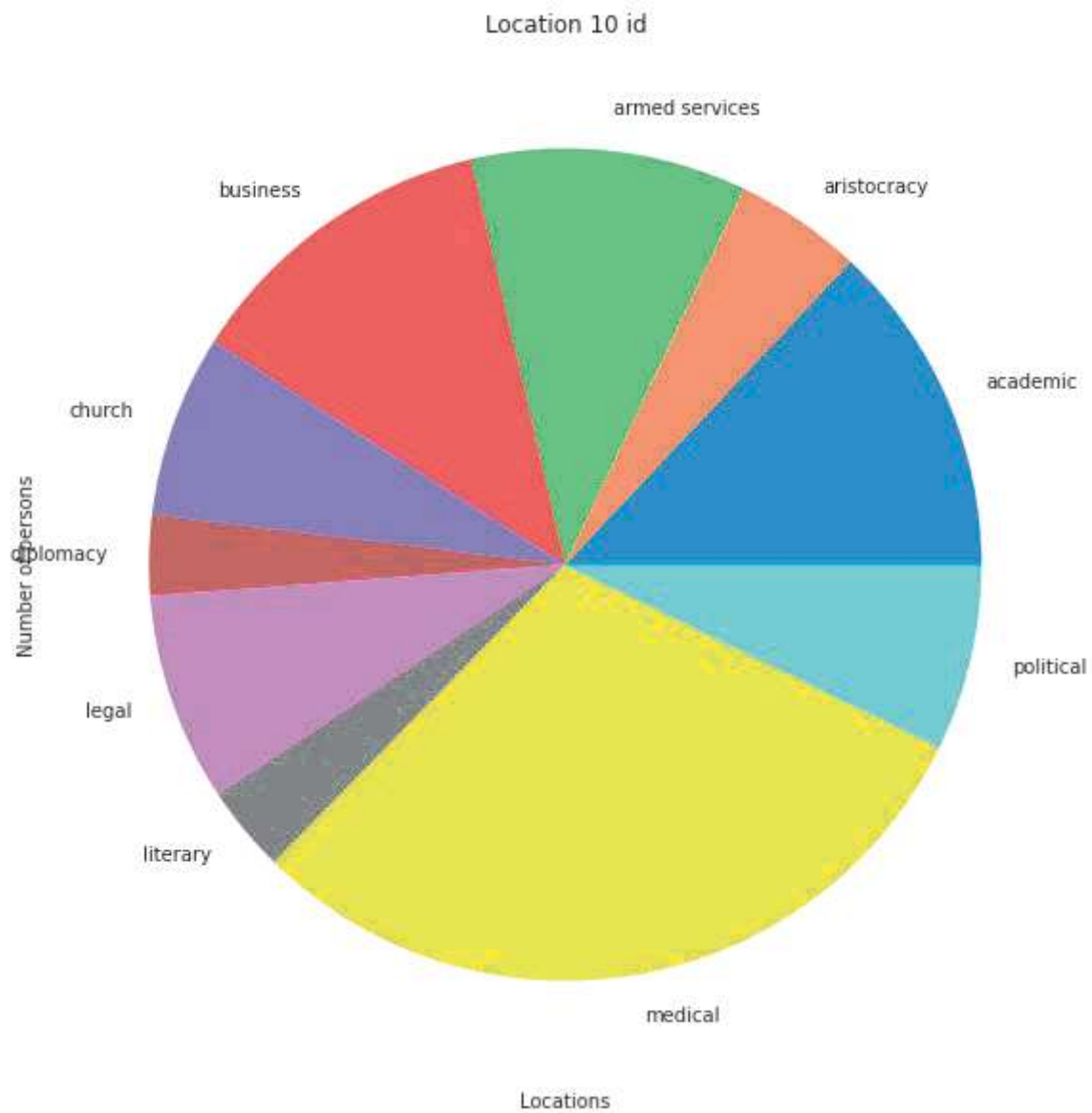
```



Occupation ID	occupation	Number
3	medical	424
7		academic
16	business	174
4	armed services	151
11	legal	114
5	political	102
6	church	100
21	aristocracy	70
1	literary	48
13	diplomacy	44
29	administrative	35

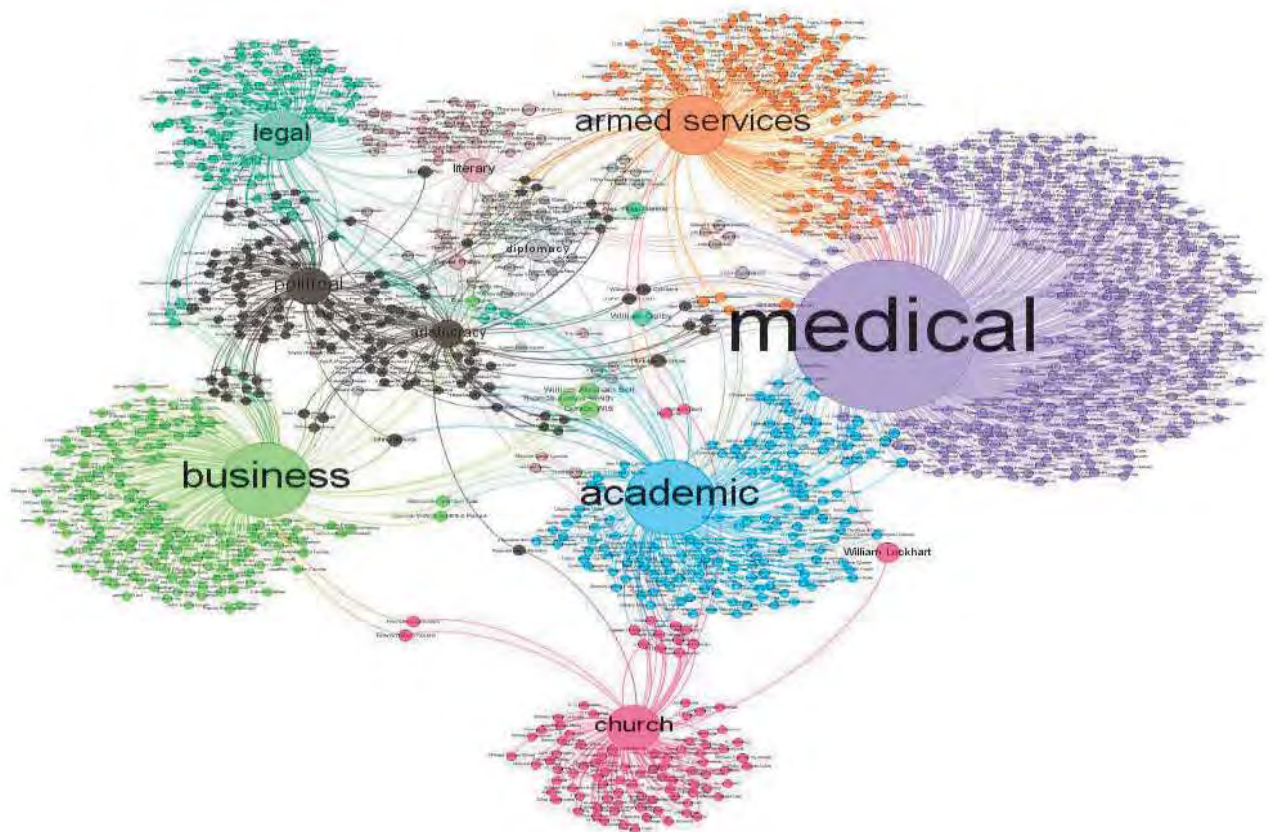
Table - The top 11 occupations

2.14 the top 10 Occupations



2.15 Some members have more than one

occupation



We can see that 'medical', 'academic' and 'armed services' together account for half of the members by occupation. We can also see that the largest three occupational categories each contain many members who have no association with any other occupational group. We can see that the medical categories contain many members who are also members of the other two principal categories ('academic' and 'armed services'). It is highly likely that these members served as conduits of communication and group cohesion amongst the three principal occupational categories.

Seven other occupations each have a range of members with literary the lowest and business the highest. All of the satellite groups relate directly to the triple centre with many members also associated with more than one other satellite occupation.

It is surprising the least networked occupation is 'church' and perhaps less so that 'business' and 'legal' are highly networked.

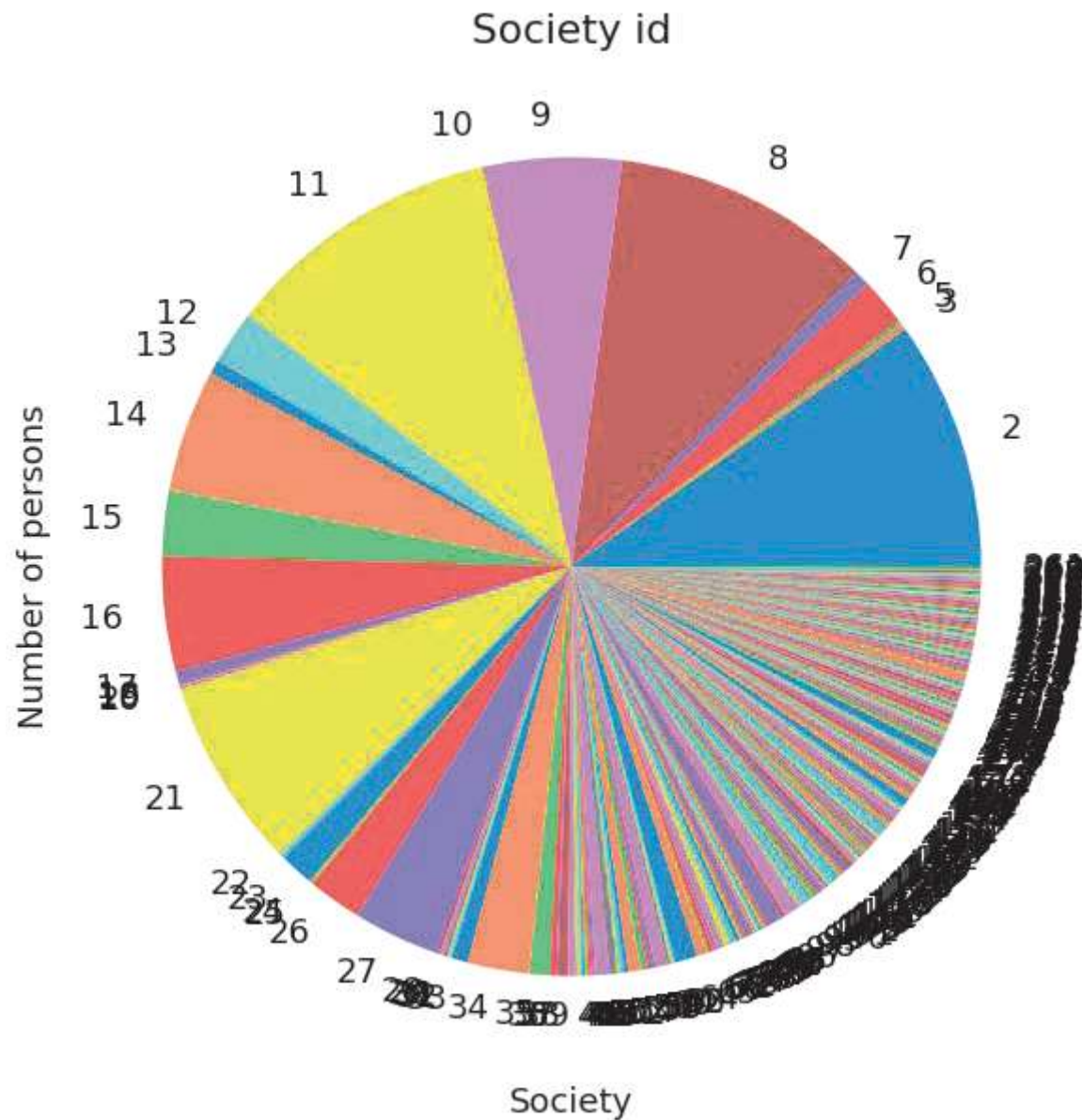
Several individuals form a web of interconnectedness between the members occupations.

2.16 Society memberships

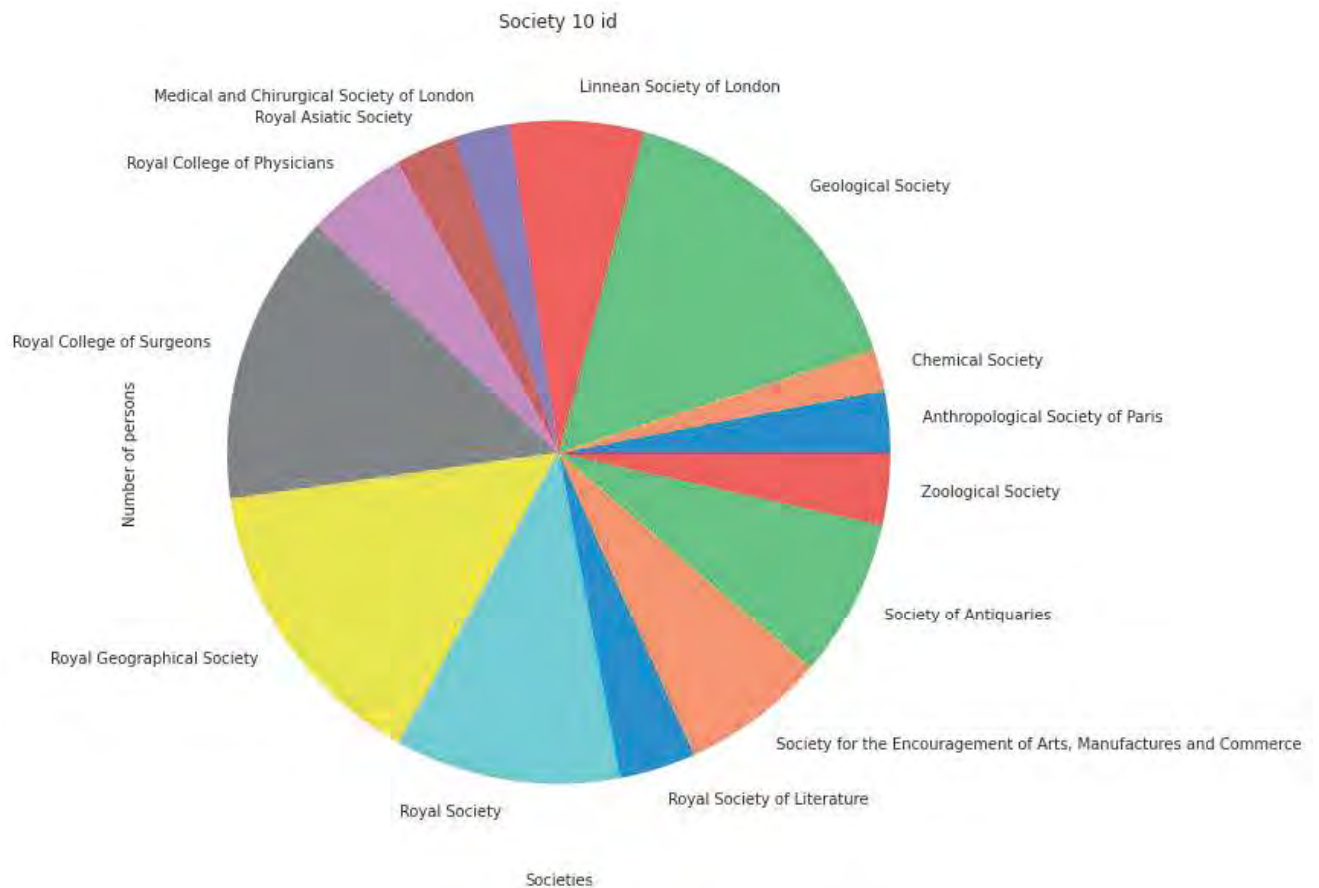
```

person_society.groupby('society_id')['person_id'].nunique().plot(kind='pie'
plt.title ("Society id")
plt.xlabel ("Society")
plt.ylabel ("Number of persons")
plt.show()

```

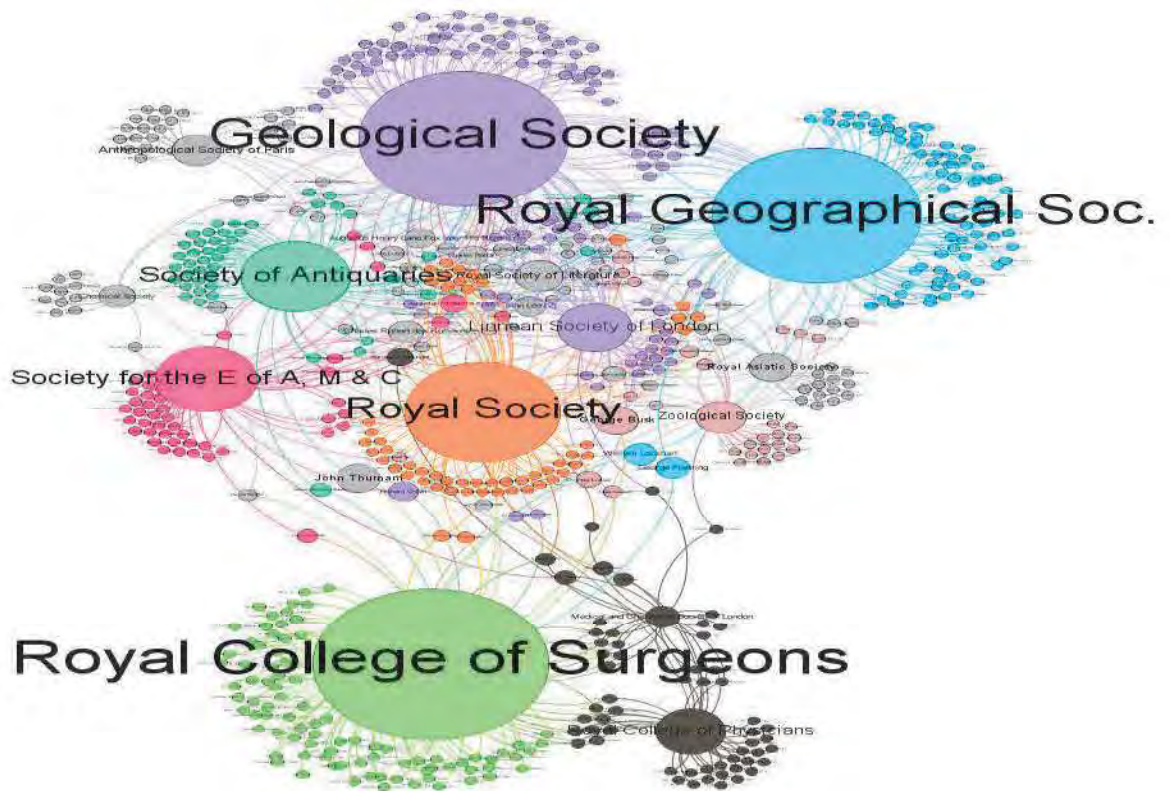


2.17 Top 10 Society memberships



2.18 Some CEDA members are members of

multiple societies



We can see that 'Geological Society' and the 'Royal Geographical Society' together account for a significant number of members by society. The 'Royal College of Surgeons', the 'Medical and Chirurgical Society' and the 'College of Physicians' form the next largest cluster of memberships of societies. These two clusters each contain many members who have no association with any other society. We can see that the medical group and the geographical group have few members in common. The 'Royal Society' and the 'Linnean Society' in the centre have between them the greatest level of networking amongst all of the societies. It is highly likely that these members served as conduits of communication and group cohesion amongst the two principal society groups.

Many other societies have a range of members all of whom are highly interconnected. All of the satellite groups relate most closely to the 'Royal Society' and the 'Linnean Society' rather than to the two larger clusters. Many members of the smaller satellite societies are also associated with more than one other satellite occupation.

It is surprising the least networked occupation is the 'Royal College of Surgeons' and perhaps less so that the 'Geological Society' and the 'Royal Geographical Society' are highly networked.

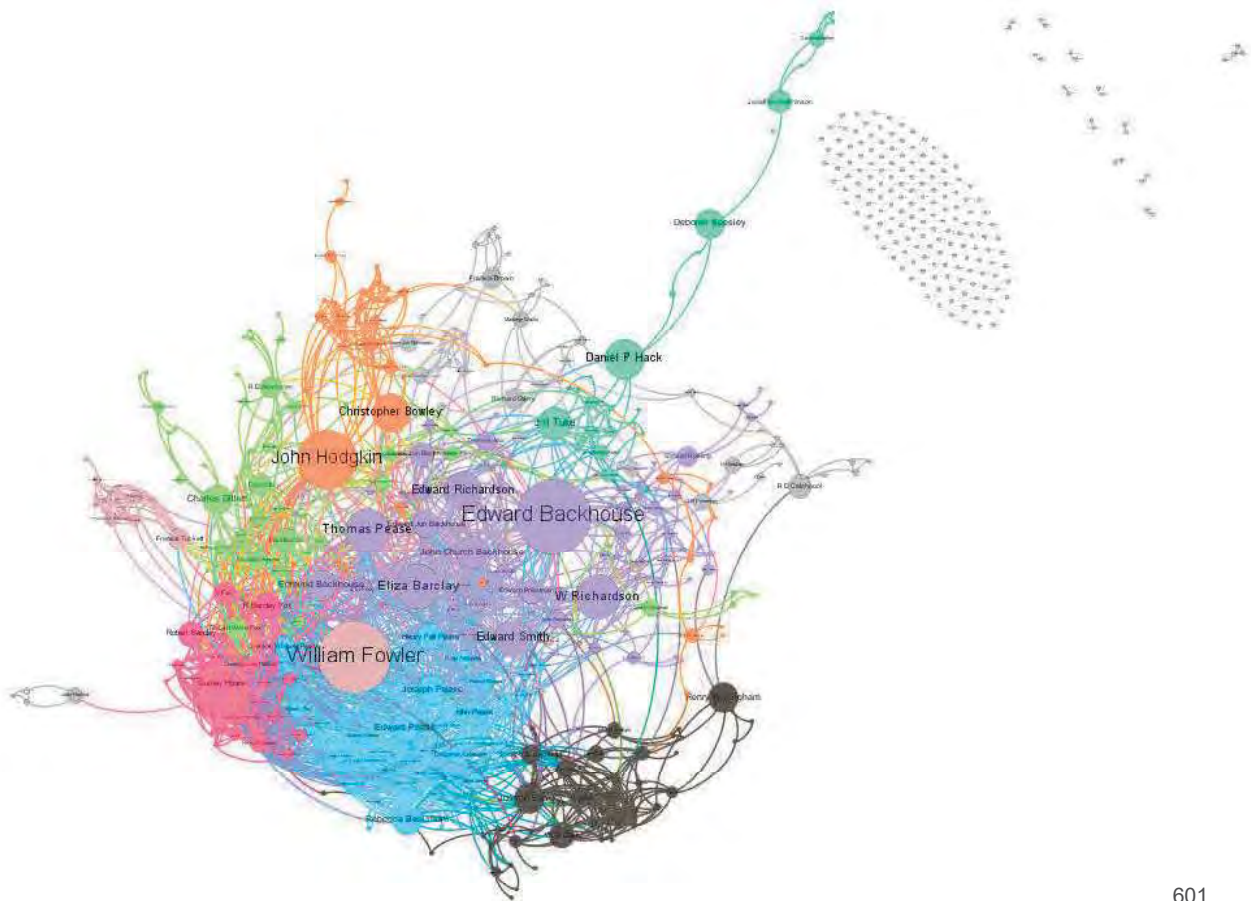
Several individuals form a web of interconnectedness between the members of societies.

2.19 All Quaker relationships

person_person

	id	relationship_type_id	person1_id	person2_id
0	1	1	23	2346
1	2	1	2264	2346
2	3	1	2265	2547
3	4	1	2545	2547
4	5	1	2546	2547
...
2094	4373	3	2494	2496
2095	4374	3	2495	2579
2096	4376	3	2867	2868
2097	4378	3	2869	2871
2098	4382	3	2503	2876

2099 rows x 4 columns



2.20 Quaker family relationships

```
person_person.groupby('relationship_type_id')['id'].nunique().plot(kind='bar')
plt.title ("Person to Person Relationships")
plt.xlabel ("Relationship Types")
plt.ylabel ("Number of persons")
plt.show()
```

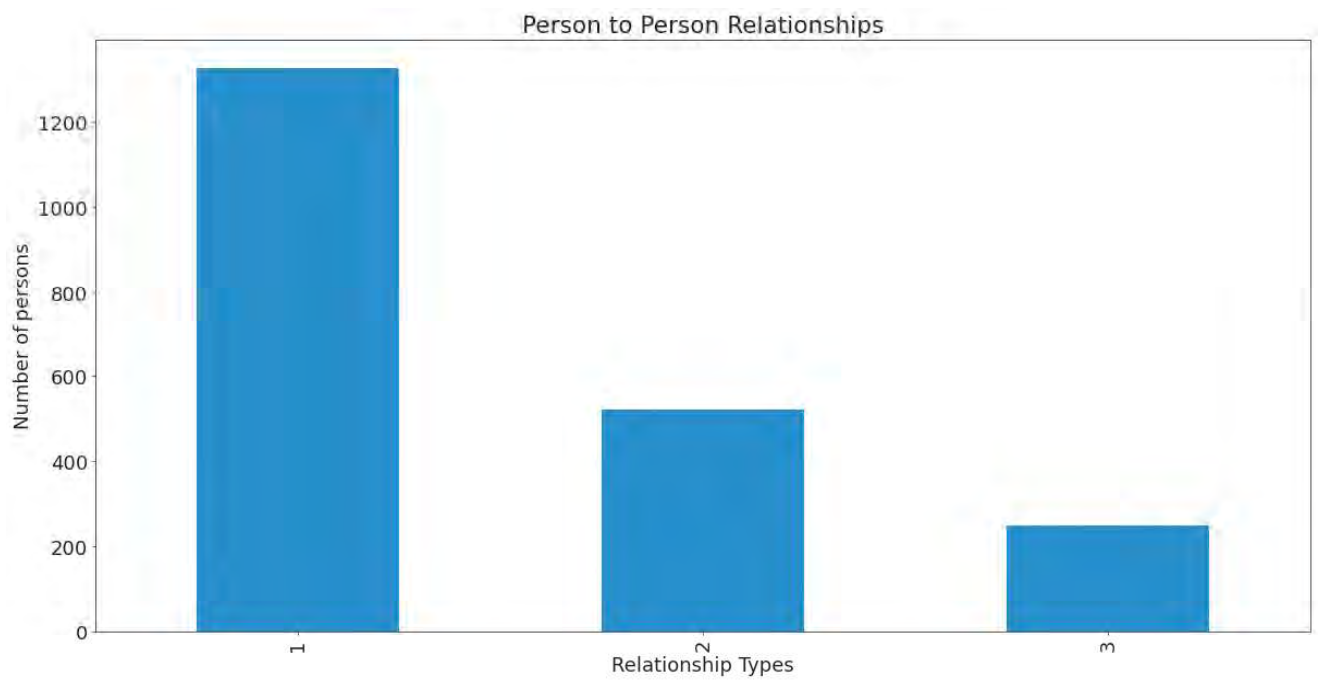
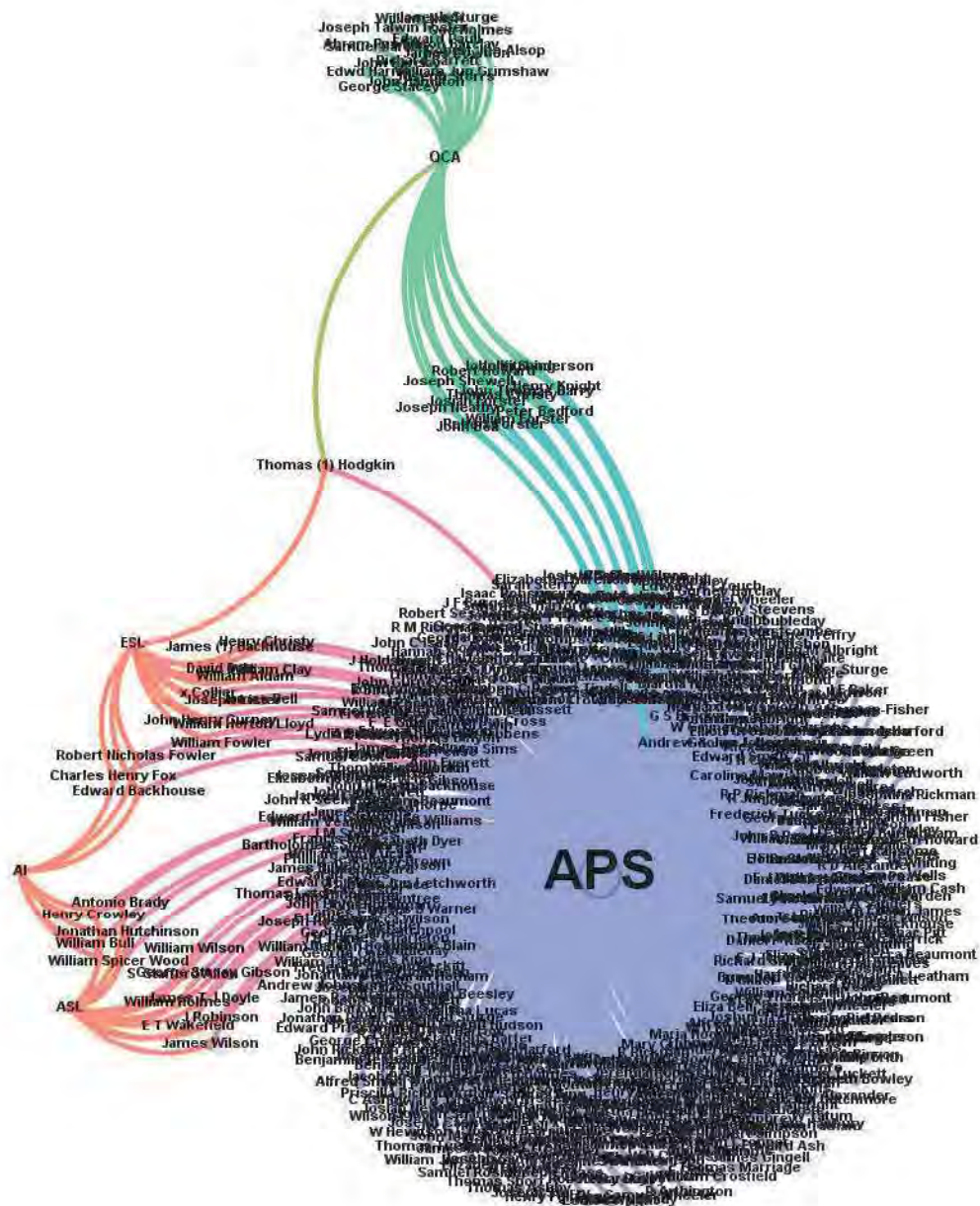


Chart above Person to person relationships (Quakers only)

Key - 1 = Distant relations, 2 = Close relations, 3 = Immediate relations

2.21 Quaker members of the CEDA



2.22 - All SQL relatable tables rendered in Gephi format

Rendering data in Gephi format

JNB uses NetworkX to generate GexF files for data to be used to produce graph files for network analysis in Gephi. Two files must be generated:

1. A file of 'Names' listing the names of all nodes to be used in a graph (These are 'persons' and all entities in related data tables, such as 'occupations'). Each row in a Names table must be unique and referenced in the 'Tuples' (or edges) table.

Note - For Names to be Gephi compliant the headers 'family_name' and 'first_names' must be combined into a single header - 'Names'.

2. A 'Tuples' (or edges) table made up of two columns of Names where the two Names are related. The first column must be headed 'Source' and the second column 'Target'.

'Names' and 'Tuples' tables can also have an 'id' column (if not then Gephi will assign one).

Note - Because 'Names', 'Source' and 'Target' are language names in Gephi capitalisation is important.

SQL views have been written to convert the data tables into Gephi standard.

```
gephi_all_names
```

	ID	Source	Target
0	1	Arthur William A Beckett	ASL
1	1	Arthur William A Beckett	London
2	1	Arthur William A Beckett	literary
3	3	Andrew Mercer Adam	ASL
4	3	Andrew Mercer Adam	armed services
...
9987	3415	x Wright	APS
9988	3416	W Wrigley	APS
9989	3417	James Yates	APS
9990	3418	John Young	APS
9991	3419	Thomas Zachary	APS

9992 rows x 3 columns

gephi_names_notceda

	ID	Source	Target
0	1	Arthur William A Beckett	London
1	1	Arthur William A Beckett	literary
2	3	Andrew Mercer Adam	armed services
3	3	Andrew Mercer Adam	country
4	3	Andrew Mercer Adam	medical
...
6093	2876	Joshua Wilson	Quaker
6094	2877	F Woodhead	Quaker
6095	2878	W Woolston	Quaker
6096	2879	Francis Wright	Quaker
6097	2880	S W Wright	Quaker

6098 rows x 3 columns

gephi_person_ceda

	ID	Source	Target
0	5	William Adam	ESL
1	7	William (1) Adams	ESL
2	8	William (2) Adams	ESL
3	12	Louis Agassiz	ESL
4	14	Augustine Aglio	ESL
...
3889	3415	x Wright	APS
3890	3416	W Wrigley	APS
3891	3417	James Yates	APS
3892	3418	John Young	APS
3893	3419	Thomas Zachary	APS

3894 rows x 3 columns

gephi_person_club

	ID	Source	Target
0	7	William (1) Adams	Athenaeum Club
1	22	Rutherford Alcock	Athenaeum Club
2	33	William Amhurst Tyssen Amhurst	Athenaeum Club
3	33	William Amhurst Tyssen Amhurst	Marlborough Club
4	33	William Amhurst Tyssen Amhurst	Carlton Club
...
318	2163	James Whishaw	Oriental Club
319	2196	S W D Williams	Athenaeum Club
320	2214	William Smith Windham	Carlton Club
321	2223	Henry Drummond Wolff	Athenaeum Club
322	2251	Ashton Yates	Athenaeum Club

323 rows x 3 columns

gephi_person_location

	ID	Source	Target
0	1	Arthur William A Beckett	London
1	3	Andrew Mercer Adam	country
2	4	H R Adam	Africa
3	6	Henry John Adams	London
4	8	William (2) Adams	London
...
2056	2255	James A Youl	London
2057	2258	Robert Younge	country
2058	2259	Arthur de Zeltner	Panama
2059	2260	x Zohrab	Armenia
2060	2260	x Zohrab	Turkey

2061 rows x 3 columns

gephi_person_occupation

	ID	Source	Target
0	1	Arthur William A Beckett	literary
1	3	Andrew Mercer Adam	medical
2	3	Andrew Mercer Adam	armed services
3	5	William Adam	political
4	8	William (2) Adams	medical
...
1878	2252	W Holt Yates	medical
1879	2253	James Yearsley	medical
1880	2254	Stephen Yeldham	medical
1881	2255	James A Youl	business
1882	2259	Arthur de Zeltner	diplomacy

1883 rows x 3 columns

gephi_person_person

	id	Source	Target
0	1	William Aldam	x Fox
1	2	William Jun Aldam	x Fox
2	3	Frederick Alexander	R D Alexander
3	4	G W Alexander	R D Alexander
4	5	Henry Alexander	R D Alexander
...
2094	4373	Alfred Waterhouse	R Waterhouse
2095	4374	Mary Waterhouse	Paul Bevan
2096	4376	Lucy Westcombe	Thomas Westcombe
2097	4378	Benjamin Wheeler	Samuel Wheeler
2098	4382	Charles Wilson	Joshua Wilson

2099 rows × 3 columns

gephi_person_religion

	ID	Source	Target
0	2233	William Spicer Wood	Quaker
1	2211	William Wilson	Quaker
2	2208	James Wilson	Quaker
3	2108	E T Wakefield	Quaker
4	1744	John Ross	Quaker
...
588	2876	Joshua Wilson	Quaker
589	2877	F Woodhead	Quaker
590	2878	W Woolston	Quaker
591	2879	Francis Wright	Quaker
592	2880	S W Wright	Quaker

593 rows × 3 columns

gephi_person_society

	ID	Source	Target
0	8	William (2) Adams	Royal College of Surgeons
1	8	William (2) Adams	Pathological Society of London
2	8	William (2) Adams	Medical Society of London
3	8	William (2) Adams	Medical and Chirurgical Society of London
4	11	William Adlam	Somersetshire Archaeological and Natural Histo...
...
1233	2245	William Cort Wright	Manchester Literary and Philosophical Society
1234	2245	William Cort Wright	Chemical Society
1235	2252	W Holt Yates	Royal College of Physicians
1236	2258	Robert Younge	York Philosophical Society
1237	2258	Robert Younge	Linnean Society of London

1238 rows x 3 columns

P7 Chapter 3 HDDT Using SQLite database standard 'views'

Index of views, dataframe info and rendering corrections

jnb_ceda_database_views

Generic code block used to set up every notebook

```
# First we call up the python packages we need to perform the analysis:

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from operator import itemgetter
import networkx as nx
from networkx.algorithms import community #This part of networkx, for commun
import nbconvert
import csv

# to add an image 

# to jump to another paragraph <a id='another_cell'></a>

# to Insert a hyperlink [https://github.com/KelvinBeerJones](https://github.

# Convert float64 to INT64;    table['column'] = table['column'].fillna(0).as
```

```
-----
ModuleNotFoundError                                Traceback (most recent call last)
Cell In[1], line 5
      3 import pandas as pd
      4 import numpy as np
----> 5 import matplotlib.pyplot as plt
      6 from operator import itemgetter
      7 import networkx as nx

ModuleNotFoundError: No module named 'matplotlib'
```

Generic code block used to make gexf file from dataframes

```
with open('vw_3_bipartite_names.csv', 'r') as nodecsv: # Open the file
    nodereader = csv.reader(nodecsv) # Read the csv
    nodes = [n for n in nodereader][1:] # Retrieve the data
    #using Python list comprhension and list slicing to remove the header row

node_names = [n[0] for n in nodes] # Get a list of only the node names

with open('vw_3_bipartite_nodes.csv', 'r') as edgecsv: # Open the file
    edgereader = csv.reader(edgecsv) # Read the csv
    edges = [tuple(e) for e in edgereader][1:] # Retrieve the data
```

```
#nodes
```



```
#edges
```

```
print(len(node_names))  
print(len(edges))
```

```
3094  
514
```

```
G = nx.Graph()  
G.add_nodes_from(node_names)  
G.add_edges_from(edges)  
print(nx.info(G))
```

```
Name:  
Type: Graph  
Number of nodes: 3869  
Number of edges: 514  
Average degree: 0.2657
```

```
nx.write_gexf(G, 'project_name')
```

3.1 Introduction and explanation

3.1.1 Introduction

In the HDDT methodology Jupyter Notebooks (JNB) are used to visualise dataframes, each of which is generated from a SQLite 'view'. JNB is used to generate charts and graphs using Pyplot and Seaborn libraries and also to generate GexF files for Gephi.

SQLite database views have been built to comprehensively 'map' the structure of the database as shown in the Entity Relationship Diagram below. Views capture:

1. Individual Name data tables - such as persons, occupations, locations, clubs and societies. Note: religion here is solely an attribute of persons (because the HDDT currently only captures Quakers). Religion would become a meaningful data table if other religious affiliations were also captured. Data tables form 'Name' tables (and Names are also known as Nodes depending on which technology is open - SQLite, JNB or Gephi).

2. Tuples tables that show many to many relationships. These are person Name(s) and their relationship to other Name(s) (occupation, location, club and society) and they are made of pairs of nodes, which combined are called a tuple in the form of 'person Name (is associated with) other Name'. Persons here are also known as 'Source' and the associated Name(s) as 'Target'.
3. Both Name and Tuple tables can have attributes attached. In Gephi attributes attached to records in a Names table will allow filtering based on Nodes whereas attributes attached to individual records in Tuples tables will allow filtering of edges based on attributes. A GexF file can contain attributes for both Names and Tuples.

Note - First letter capitalisation in the HDDT must be followed. The Gephi dictionary requires Name, Source and Target to be in this form.

The process:

1. Devise a set of comprehensive database views (Using DBeaver).
2. Export the views as csv files to the container 'jnd_ceda_database_sql_views'.
3. This container is also a GitHub repo to enable version control.
4. This Jupyter Notebook is located here (all resources necessary for a JNB must be in the same container).
5. Use JNB to make a dataframe for each csv file and display the first 10 records.
6. Check the dataframe info to ensure that no tables have columns rendered as float64. (Apply the method: `table['column'] = table['column'].fillna(0).astype(np.int64)` to covert float64 to int64 if necessary.
7. Make subsets of dataframes to slice the data. (Use SQLite database select queries to make INNER and LEFT joins between tables)

3.1.2 Explanation

This workbook resides in a GitHub container facilitating version control. Each time this workbook is amended a record is made in the corresponding GitHub repo.

Gephi requires a Names file (generated by Networkx), and this comprises of all Nodes irrespective of which side of an EDGE they will later be attached to. In the Edges file in Gephi the Names now become Nodes paired as Tuples (Source and Target) and Gephi infers an Edge between them. Person names and bipartite group names are variously called - Names, Nodes, Source and Target depending on where they are being used.

Names files in Gephi. Names can have attributes and these can be used to style Nodes.

Tuples in Gephi format = First column must be "Source", second column must be "Target", additional columns are 'attributes' of each tuple. (Note: attributes will be used in Gephi to style edges and not nodes.)

All dataframes consist of data types of OBJECT and INT64 only. CSV sheets occasionally render columns that render as FLOAT64. Where this occurs a fix is applied immediately after the `pd.read_csv` command.

3.2 To make a new project

Make a new project repo in Github.

1 Clone the new project repo to a new container in the HDDT workspace.

2 Create a JNB notebook in the new container workspace.

3 Copy selected csv files from this container to the new container.

4 In the new JNB use routines to validate selected data.-

```
pd.read_csv,  
df.iloc [0:10]  
df.info ()
```

5 Use this routine to correct float64 columns. `df['column'] = df['column'].astype(int)`

6 Slice data as needed and make a new df of the subset data.

7 Use pandas to make graphs of the data.

8 If wanted, use the routine `df.to_csv ('vw_hddt_newdataframe.csv')` to put a

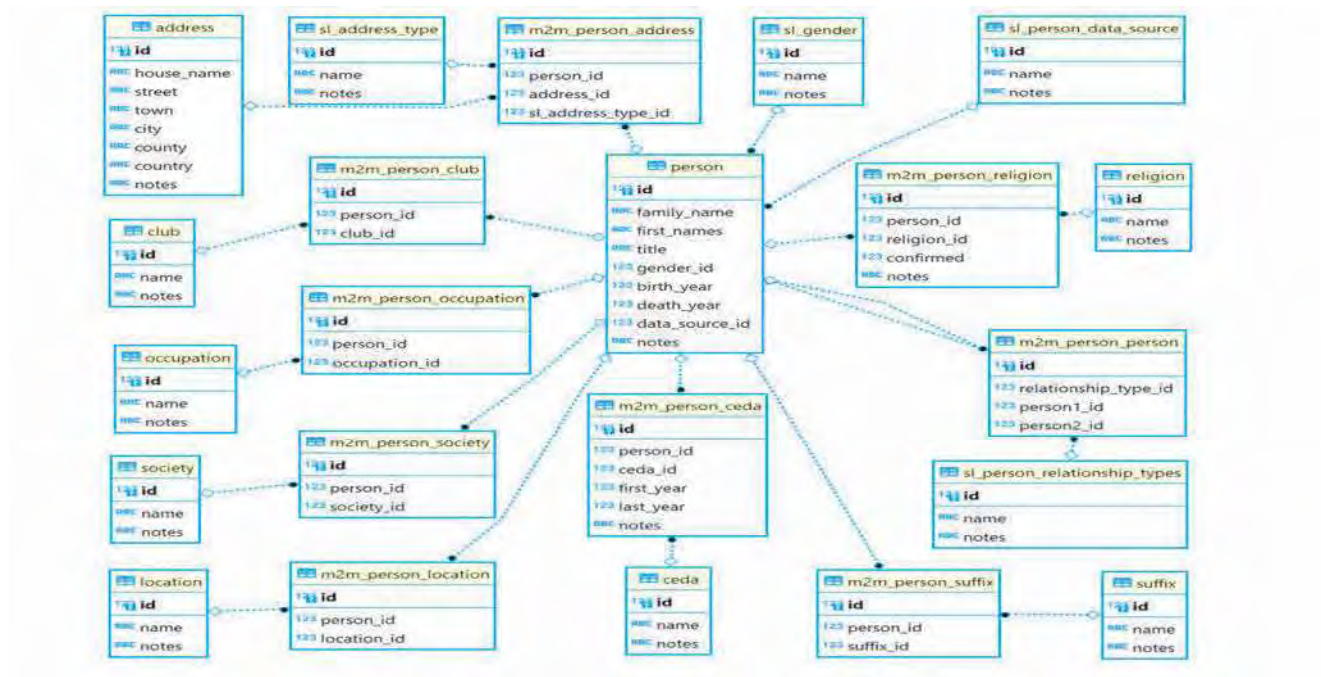
9 If wanted, make a GexF file (in the new workspace container) to use Gephi

10 Don't forget to use VSC to update Github for version control if changes are made
for this jnb for version control and to make latest version available to all

Links to sections in this workbook

Table
Person Table
Person Attributes
Person names and other nodes combined
Other Nodes
Bigraph aa tuples
Religion tuples
Location tuples
Occupation tuples
Society tuples
Club tuples
CEDA Name attributes
CEDA tuples
CEDA tuples attributes
Quakers
Quaker immediate
Quaker close
Quaker distant
Quaker CEDA tuples

3.3 Entity Relationship Diagram



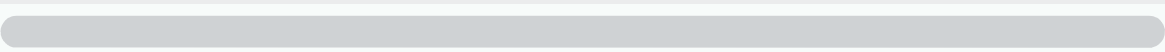
3.4 Person table (3094 records)

Dataframe

```
person_table = pd.read_csv ('vw_hddt_person_table.csv')
person_table ['gender_id'] = person_table ['gender_id'].fillna(0).astype(np
person_table ['birth_year'] = person_table ['birth_year'].fillna(0).astype(i
person_table ['death_year'] = person_table ['death_year'].fillna(0).astype(np
```

```
person_table.iloc [0:10]
```

	Name	title	gender_id	birth_year	death_year	data_source_id	
0	Arthur William A Beckett	NaN	1	1844	1909	1	S J S.A Ja
1	Andrew Mercer Adam	NaN	1	0	0	1	Linc
2	H R Adam	NaN	1	0	0	1	Old c V
3	William Adam	NaN	1	0	0	1	
4	Henry John Adams	NaN	1	0	0	1	14 T Sq
5	William (1) Adams	NaN	1	0	0	1	
6	William (2) Adams	NaN	1	1820	1900	1	5 H Ca [180
7	William Adlam	NaN	1	0	0	1	Stre [1863
8	Louis Agassiz	NaN	1	1807	1873	1	C Car
9	Anastasius Agathides	NaN	1	1805	1881	1	28 Wes [A3]



```
person_table.info ()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3094 entries, 0 to 3093
Data columns (total 7 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Name                   3094 non-null   object
1   title                  776 non-null    object
2   gender_id              3094 non-null   int64
3   birth_year             3094 non-null   int64
4   death_year             3094 non-null   int64
5   data_source_id         3094 non-null   int64
6   notes                  1770 non-null   object
dtypes: int64(4), object(3)
memory usage: 169.3+ KB
```

3.5 Person Names (3094 records)

Datatable

```
person_name = pd.read_csv ('vw_hddt_person_name.csv')
```

```
person_name.iloc [0:10]
```

	Name
0	Arthur William A Beckett
1	Andrew Mercer Adam
2	H R Adam
3	William Adam
4	Henry John Adams
5	William (1) Adams
6	William (2) Adams
7	William Adlam
8	Louis Agassiz
9	Anastasius Agathides

```
person_name.info ()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3094 entries, 0 to 3093
Data columns (total 1 columns):
#   Column   Non-Null Count  Dtype
---  -
0    Name     3094 non-null   object
dtypes: object(1)
memory usage: 24.3+ KB
```

3.6 Persons with attributes (Names file) (3094 records)

Datatable

```
person_attributes = pd.read_csv ('vw_hddt_person_attributes_religion.csv')
person_attributes ['religion_1_quaker'] = person_attributes ['religion_1_quaker']
person_attributes ['birth_year'] = person_attributes ['birth_year'].fillna(0)
person_attributes['death_year'] = person_attributes['death_year'].fillna(0)
```

```
person_attributes.iloc [0:10]
```

	Name	birth_year	death_year	religion_1_quaker
0	Arthur William A Beckett	1844	1909	0
1	Andrew Mercer Adam	0	0	0
2	H R Adam	0	0	0
3	William Adam	0	0	0
4	Henry John Adams	0	0	0
5	William (1) Adams	0	0	0
6	William (2) Adams	1820	1900	0
7	William Adlam	0	0	0
8	Louis Agassiz	1807	1873	0
9	Anastasius Agathides	1805	1881	0


```
person_attributes.info ()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3094 entries, 0 to 3093
Data columns (total 4 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Name                   3094 non-null   object
1   birth_year             3094 non-null   int64
2   death_year             3094 non-null   int64
3   religion_1_quaker      3094 non-null   int64
dtypes: int64(3), object(1)
memory usage: 96.8+ KB
```

3.7 All Names (Nodes) (3608 records)

Dataframe

All person names (3095) and bipartite nodes (514) in Gephi 'Names' format. Can be used with a 'tuples' file to generate a GexF file for Gephi where all possible nodes would appear on the visualisation, including nodes with no associated tuple.

```
all_names_and_nodes = pd.read_csv('vw_hddt_all_names_and_nodes.csv')
```

```
all_names_and_nodes.loc [0:10]
```

	Name
0	Joseph Storrs
1	A Mackintosh Shaw
2	A de Fullner
3	A , jun Ramsay
4	A A Stewart
5	A Ambrose
6	A B Stark
7	A B Wright
8	A Bell
9	A C Brebner
10	A Crowley

```
all_names_and_nodes.info ()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3608 entries, 0 to 3607
Data columns (total 1 columns):
#   Column  Non-Null Count  Dtype
---  -
0    Name    3608 non-null    object
dtypes: object(1)
memory usage: 28.3+ KB
```

3.8 All bipartite Names (514 records)

Dataframe

Bigraph nodes in Gephi Name format. This is a subset of 'all_names_and_nodes'

```
bigraph_nodes = pd.read_csv ('vw_hddt_bigraph_nodes.csv')
```

```
bigraph_nodes.iloc [0:10]
```

	Name
0	AI
1	APS
2	ASL
3	Aberdeen Horticultural Society
4	Academia Quirurgia of Madrid
5	Academie Hongroise de Pest
6	Academy of Anatolia
7	Academy of Medicine and Surgery of Madrid and ...
8	Academy of Natural Sciences Philadelphia
9	Academy of Natural Sciences of Spain

```
bigraph_nodes.info ()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 514 entries, 0 to 513
Data columns (total 1 columns):
#   Column  Non-Null Count  Dtype
---  ---
0    Name    514 non-null      object
dtypes: object(1)
memory usage: 4.1+ KB
```

3.9 All Names (Nodes) as Tuples (9989 records)

dataframe

All tuples from the HDDT in Gephi format. (There are 9991 edges between the 3095 persons and the 514 Bipartite nodes.)

```
bigraph_all_tuples = pd.read_csv ('vw_hddt_all_bigraph_tuples.csv')
```

```
bigraph_all_tuples.iloc [0:10]
```

	Source	Target
0	Joseph Storrs	QCA
1	Joseph Storrs	Quaker
2	A Mackintosh Shaw	ASL
3	A Mackintosh Shaw	country
4	A de Fullner	AI
5	A , jun Ramsay	AI
6	A , jun Ramsay	ASL
7	A , jun Ramsay	Geological Society
8	A , jun Ramsay	London
9	A A Stewart	ASL

```
bigraph_all_tuples.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9989 entries, 0 to 9988
Data columns (total 2 columns):
#   Column  Non-Null Count  Dtype
---  ---
0    Source  9989 non-null     object
1    Target  9989 non-null     object
dtypes: object(2)
memory usage: 156.2+ KB
```

3.10 Religion tuples (592 records)

Dataframe

Quakers

```
religion_tuples = pd.read_csv ('vw_hddt_religion_tuples.csv')
```

```
religion_tuples.iloc [0:10]
```

	Source	Target
0	William Spicer Wood	Quaker
1	William Wilson	Quaker
2	James Wilson	Quaker
3	E T Wakefield	Quaker
4	John Ross	Quaker
5	J Robinson	Quaker
6	William Horton Lloyd	Quaker
7	Joseph Lister	Quaker
8	Jonathan Hutchinson	Quaker
9	William Holmes	Quaker

```
religion_tuples.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 592 entries, 0 to 591
Data columns (total 2 columns):
#   Column  Non-Null Count  Dtype
---  -
0    Source  592 non-null      object
1    Target  592 non-null      object
dtypes: object(2)
memory usage: 9.4+ KB
```

3.11 Location tuples (2061 records)

Dataframe

Location (UK but not London)

```
location_tuples = pd.read_csv ('vw_hddt_location_tuples.csv')
```

```
location_tuples.iloc [0:10]
```

	Source	Target
0	Arthur William A Beckett	London
1	Andrew Mercer Adam	country
2	H R Adam	Africa
3	Henry John Adams	London
4	William (2) Adams	London
5	William Adlam	country
6	Louis Agassiz	America
7	Anastasius Agathides	London
8	Joseph Agnew	Scotland
9	William Francis Harrison Ainsworth	London

```
location_tuples.info ()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2061 entries, 0 to 2060
Data columns (total 2 columns):
#   Column  Non-Null Count  Dtype
---  -
0    Source  2061 non-null    object
1    Target  2061 non-null    object
dtypes: object(2)
memory usage: 32.3+ KB
```

3.12 Occupation tuples (1883 records)

Dataframe

Occupations

```
occupation_tuples = pd.read_csv ('vw_hddt_occupation_tuples.csv')
```

```
occupation_tuples.iloc [0:10]
```

	Source	Target
0	Arthur William A Beckett	literary
1	Andrew Mercer Adam	medical
2	Andrew Mercer Adam	armed services
3	William Adam	political
4	William (2) Adams	medical
5	Louis Agassiz	academic
6	Louis Agassiz	biologist
7	Louis Agassiz	geologist
8	Anastasius Agathides	academic
9	Augustine Aglio	artist

```
occupation_tuples.info ()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1883 entries, 0 to 1882
Data columns (total 2 columns):
#   Column  Non-Null Count  Dtype
---  -
0    Source  1883 non-null    object
1    Target  1883 non-null    object
dtypes: object(2)
memory usage: 29.5+ KB
```

3.13 Society tuples (1238 records)

Dataframe

Society memberships

```
society_tuples = pd.read_csv ('vw_hddt_society_tuples.csv')
```

```
society_tuples.iloc [0:10]
```

	Source	Target
0	William (2) Adams	Royal College of Surgeons
1	William (2) Adams	Pathological Society of London
2	William (2) Adams	Medical Society of London
3	William (2) Adams	Medical and Chirurgical Society of London
4	William Adlam	Somersetshire Archaeological and Natural Histo...
5	William Francis Harrison Ainsworth	Royal Geographical Society
6	William Francis Harrison Ainsworth	Society of Antiquaries
7	William Francis Harrison Ainsworth	Syro Egyptian Society
8	William Francis Harrison Ainsworth	Geological Society
9	William Baird Airston	Royal College of Surgeons

```
society_tuples.info ()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1238 entries, 0 to 1237
Data columns (total 2 columns):
#   Column  Non-Null Count  Dtype
---  ------  -
0    Source   1238 non-null    object
1    Target   1238 non-null    object
dtypes: object(2)
memory usage: 19.5+ KB
```

3.14 Club tuples (323 records)

Dataframe

Club memberships

```
club_tuples = pd.read_csv ('vw_hddt_club_tuples.csv')
```



```
club_tuples.iloc [0:10]
```

	Source	Target
0	William (1) Adams	Athenaeum Club
1	Rutherford Alcock	Athenaeum Club
2	William Amhurst Tyssen Amhurst	Athenaeum Club
3	William Amhurst Tyssen Amhurst	Marlborough Club
4	William Amhurst Tyssen Amhurst	Carlton Club
5	William Arbuthnot	Oriental Club
6	Richard Edward Arden	National Club
7	Richard Edward Arden	Junior Athenaeum Club
8	William Armstrong	Athenaeum Club
9	William Henry Ashurst	Reform Club

```
club_tuples.info ()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 323 entries, 0 to 322
Data columns (total 2 columns):
#   Column  Non-Null Count  Dtype
---  ------  -
0    Source   323 non-null     object
1    Target   323 non-null     object
dtypes: object(2)
memory usage: 5.2+ KB
```

3.15 CEDA tuples (3892 records)

Dataframe

Tuples in Gephi format to graph the memberships of CEDA

```
ceda_tuples = pd.read_csv('vw_hddt_ceda_tuples.csv')
```

```
ceda_tuples.iloc [0:10]
```

	Source	Target
0	William Adam	ESL
1	William (1) Adams	ESL
2	William (2) Adams	ESL
3	Louis Agassiz	ESL
4	Augustine Aglio	ESL
5	William Francis Harrison Ainsworth	ESL
6	Alexander Muirhead Aitken	ESL
7	Rutherford Alcock	ESL
8	William Aldam	ESL
9	William Allen	ESL

```
ceda_tuples.info ()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3892 entries, 0 to 3891
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype  
---  -
 0   Source  3892 non-null    object  
 1   Target  3892 non-null    object  
dtypes: object(2)
memory usage: 60.9+ KB
```

3.16 Quaker Committee on the Aborigines (QCA)

```
qca = ceda_tuples[ceda_tuples["Target"] == "QCA"]
```

```
qca.iloc [0:10]
```

	Source	Target
2732	Thomas (1) Hodgkin	QCA
2998	James Bowden	QCA
2999	William Nash	QCA
3000	Joseph Sturge	QCA
3001	William Jun Grimshaw	QCA
3003	Henry Knight	QCA
3004	Edward Paull	QCA
3005	Robert Jun Alsop	QCA
3006	Abram Rawlinson Barclay	QCA
3007	John Barclay	QCA

```
qca.to_csv ('vw_hddt_ceda_qca.csv')
```

3.17 Aborigines Protection Society (APS)

```
aps = ceda_tuples[ceda_tuples["Target"] == "APS"]
```

```
aps.iloc [0:10]
```

	Source	Target
2692	William Aldam	APS
2693	Samuel C Baker	APS
2694	James Bell	APS
2695	John Bell (2)	APS
2696	John Brown	APS
2697	Henry Christy	APS
2698	Thomas junior Christy	APS
2699	William Clay	APS
2700	Richard King	APS
2701	John James Sturz	APS

```
aps.to_csv ('vw_hddt_ceda_aps.csv')
```

3.18 Ethnological Society of London (ESL)

```
esl = ceda_tuples[ceda_tuples["Target"] == "ESL"]
```

```
esl.iloc [0:10]
```

	Source	Target
0	William Adam	ESL
1	William (1) Adams	ESL
2	William (2) Adams	ESL
3	Louis Agassiz	ESL
4	Augustine Aglio	ESL
5	William Francis Harrison Ainsworth	ESL
6	Alexander Muirhead Aitken	ESL
7	Rutherford Alcock	ESL
8	William Aldam	ESL
9	William Allen	ESL

```
esl.to_csv ('vw_hddt_ceda_esl.csv')
```

3.19 Anthropological Society of London (ASL)

```
asl = ceda_tuples[ceda_tuples["Target"] == "ESL"]
```

```
asl.iloc [0:10]
```

	Source	Target
0	William Adam	ESL
1	William (1) Adams	ESL
2	William (2) Adams	ESL
3	Louis Agassiz	ESL
4	Augustine Aglio	ESL
5	William Francis Harrison Ainsworth	ESL
6	Alexander Muirhead Aitken	ESL
7	Rutherford Alcock	ESL
8	William Aldam	ESL
9	William Allen	ESL

```
asl.to_csv ('vw_hddt_ceda_asl.csv')
```

3.20 Anthropological Institute (AI)

```
ai = ceda_tuples[ceda_tuples["Target"] == "ESL"]
```

```
ai.iloc [0:10]
```

	Source	Target
0	William Adam	ESL
1	William (1) Adams	ESL
2	William (2) Adams	ESL
3	Louis Agassiz	ESL
4	Augustine Aglio	ESL
5	William Francis Harrison Ainsworth	ESL
6	Alexander Muirhead Aitken	ESL
7	Rutherford Alcock	ESL
8	William Aldam	ESL
9	William Allen	ESL

```
ai.to_csv ('vw_hddt_ceda_ai.csv')
```

3.21CEDA Name with attributes (3892 records)

Dataframe

Datatable of all people and their memberships of CEDA (some people are in more than one). Attaches attributes to Nodes in Gephi. (Note: records = greater than 3095 persons due to multiple memberships). Gephi will disregard duplicate Names (but not tuples).

```
ceda_name_attributes = pd.read_csv ('vw_hddt_ceda_name_attributes.csv')
ceda_name_attributes ['quaker'] = ceda_name_attributes ['quaker'].fillna(0)
```

```
ceda_name_attributes.iloc [0:10]
```

	Name	quaker	first_year	last_year	birth_year	death_year
0	William Adam	0	1844	1844	NaN	NaN
1	William (1) Adams	0	1844	1844	NaN	NaN
2	William (2) Adams	0	1858	1871	1,820	1,900
3	Louis Agassiz	0	1860	1871	1,807	1,873
4	Augustine Aglio	0	1843	1845	1,777	1,857
5	William Francis Harrison Ainsworth	0	1856	1860	1,807	1,896
6	Alexander Muirhead Aitken	0	1864	1871	NaN	NaN
7	Rutherford Alcock	0	1862	1871	1,809	1,897
8	William Aldam	1	1844	1848	1,813	1,890
9	William Allen	0	1858	1858	NaN	NaN

```
ceda_name_attributes.info ()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3892 entries, 0 to 3891
Data columns (total 6 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Name             3892 non-null   object
1   quaker           3892 non-null   int64
2   first_year       3892 non-null   int64
3   last_year        3892 non-null   int64
4   birth_year       1528 non-null   object
5   death_year       1639 non-null   object
dtypes: int64(3), object(3)
memory usage: 182.6+ KB
```

3.22 CEDA tuples with attributes (3892 records)

Dataframe

CEDA 'tuples with attributes' attaches attributes to edges in Gephi.


```
ceda_tuples_attributes = pd.read_csv ('vw_hddt_ceda_tuples_attributes.csv')
```

```
ceda_tuples_attributes.iloc [0:10]
```

	Source	Target	first_year	last_year	birth_year	death_year
0	William Adam	ESL	1844	1844	NaN	NaN
1	William (1) Adams	ESL	1844	1844	NaN	NaN
2	William (2) Adams	ESL	1858	1871	1,820	1,900
3	Louis Agassiz	ESL	1860	1871	1,807	1,873
4	Augustine Aglio	ESL	1843	1845	1,777	1,857
5	William Francis Harrison Ainsworth	ESL	1856	1860	1,807	1,896
6	Alexander Muirhead Aitken	ESL	1864	1871	NaN	NaN
7	Rutherford Alcock	ESL	1862	1871	1,809	1,897
8	William Aldam	ESL	1844	1848	1,813	1,890
9	William Allen	ESL	1858	1858	NaN	NaN

```
ceda_tuples_attributes.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3892 entries, 0 to 3891
Data columns (total 6 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Source           3892 non-null   object
1   Target           3892 non-null   object
2   first_year       3892 non-null   int64
3   last_year        3892 non-null   int64
4   birth_year       1528 non-null   object
5   death_year       1639 non-null   object
dtypes: int64(2), object(4)
memory usage: 182.6+ KB
```

3.23 Quakers (592 records)

Dataframe

```
quakers = pd.read_csv ('vw_hddt_quakers.csv')
quakers ['birth_year'] = quakers ['birth_year'].fillna(0).astype(np.int64)
quakers ['death_year'] = quakers ['death_year'].fillna(0).astype(np.int64)
```

```
quakers.iloc [0:10]
```

	Name	birth_year	death_year
0	William Aldam	1813	1890
1	S Stafford Allen	1840	1870
2	Edward Backhouse	1808	1879
3	James (1) Backhouse	1794	1869
4	James Bell	1818	1872
5	Antonio Brady	1811	1881
6	William Bull	1828	1902
7	Charles Buxton	1823	1871
8	Henry Christy	1810	1865
9	William Clay	1791	1869

```
quakers.info ()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 592 entries, 0 to 591
Data columns (total 3 columns):
#   Column      Non-Null Count  Dtype
---  -
0   Name        592 non-null   object
1   birth_year  592 non-null   int64
2   death_year  592 non-null   int64
dtypes: int64(2), object(1)
memory usage: 14.0+ KB
```

3.24 Quaker family relationships (2086 records)

Dataframe

```
person_relationships = pd.read_csv ('vw_hddt_person1_person2.csv')
```

```
person_relationships.iloc [0:10]
```

	Source	Target	relationship_type_id
0	William Aldam	x Fox	1
1	William Jun Aldam	x Fox	1
2	Frederick Alexander	R D Alexander	1
3	G W Alexander	R D Alexander	1
4	Henry Alexander	R D Alexander	1
5	R D Alexander	John M Candler	1
6	Thomas Allis	James Jun Backhouse	1
7	Thomas Allis	Francis Brown	1
8	Thomas Allis	Septimus Warner	1
9	R Arthington	J Gurney Barclay	1

```
person_relationships.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2086 entries, 0 to 2085
Data columns (total 3 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Source                2086 non-null  object
1   Target                2086 non-null  object
2   relationship_type_id  2086 non-null  int64
dtypes: int64(1), object(2)
memory usage: 49.0+ KB
```

3.25 Quaker immediate relationships (246 records)

Datatable

```
quaker_immediate_relationships = pd.read_csv ('vw_hddt_person_person_immediate')
```

```
quaker_immediate_relationships.iloc [0:10]
```

	'Source'	Target	immediate
0	Source	John M Albright	3
1	Source	Rachel Albright	3
2	Source	William Albright	3
3	Source	John M Albright	3
4	Source	William Albright	3
5	Source	John M Albright	3
6	Source	William Jun Aldam	3
7	Source	G W Alexander	3
8	Source	Henry Alexander	3
9	Source	Henry Alexander	3

```
quaker_immediate_relationships.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 246 entries, 0 to 245
Data columns (total 3 columns):
#   Column      Non-Null Count  Dtype
---  -
0   'Source'    246 non-null   object
1   Target      246 non-null   object
2   immediate   246 non-null   int64
dtypes: int64(1), object(2)
memory usage: 5.9+ KB
```

3.26 Quakers close relationships (519 records)

Datatable

```
quaker_close_relationships = pd.read_csv ('vw_hddt_person_person_close.csv')
```

```
quaker_close_relationships.iloc [0:10]
```

	'Source'	Target	close
0	Source	Christopher Bowley	2
1	Source	Robert Charleton	2
2	Source	Frederick H Fox	2
3	Source	Thomas Maw	2
4	Source	William Norton	2
5	Source	Algernon Peckover	2
6	Source	Cornelius Hanbury	2
7	Source	Edward Beck	2
8	Source	Cornelius Hanbury	2
9	Source	Martha Lucas	2

```
quaker_close_relationships.info ()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 519 entries, 0 to 518
Data columns (total 3 columns):
#   Column      Non-Null Count  Dtype
---  -
0   'Source'    519 non-null   object
1   Target      519 non-null   object
2   close       519 non-null   int64
dtypes: int64(1), object(2)
memory usage: 12.3+ KB
```

3.27 Quaker distant relationships (1321 records)

Datatable

```
quaker_distant_relationships = pd.read_csv ('vw_hddt_person_person_distant.csv')
```

```
quaker_distant_relationships.iloc [0:10]
```

	'Source'	Target	distant
0	Source	x Fox	1
1	Source	x Fox	1
2	Source	R D Alexander	1
3	Source	R D Alexander	1
4	Source	R D Alexander	1
5	Source	John M Candler	1
6	Source	James Jun Backhouse	1
7	Source	Francis Brown	1
8	Source	Septimus Warner	1
9	Source	J Gurney Barclay	1

```
quaker_distant_relationships.info ()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1321 entries, 0 to 1320
Data columns (total 3 columns):
#   Column      Non-Null Count  Dtype
---  -
0   'Source'    1321 non-null  object
1   Target      1321 non-null  object
2   distant     1321 non-null  int64
dtypes: int64(1), object(2)
memory usage: 31.1+ KB
```

3.28 Quaker CEDA membership (tuples) (643 records)

Dataframe

```
quakers_ceda_tuples = pd.read_csv ('vw_hddt_quakers_ceda_tuples.csv')
```

```
quakers_ceda_tuples.iloc [0:10]
```

	Source	Target	religion_name	first_year	last_year
0	William Spicer Wood	APS	Quaker	1864	1867
1	William Spicer Wood	ASL	Quaker	1863	1871
2	William Spicer Wood	AI	Quaker	1863	1871
3	William Wilson	APS	Quaker	1838	1865
4	William Wilson	ASL	Quaker	1865	1866
5	James Wilson	APS	Quaker	1862	1867
6	James Wilson	ASL	Quaker	1865	1865
7	E T Wakefield	APS	Quaker	1853	1864
8	E T Wakefield	ASL	Quaker	1865	1868
9	John Ross	APS	Quaker	1839	1852

```
quakers_ceda_tuples.info ()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 643 entries, 0 to 642
Data columns (total 5 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Source          643 non-null   object
1   Target          643 non-null   object
2   religion_name    643 non-null   object
3   first_year      643 non-null   int64
4   last_year       643 non-null   int64
dtypes: int64(2), object(3)
memory usage: 25.2+ KB
```

3.29 Quakers in the QCA

```
quakers_qca = quakers_ceda_tuples[quakers_ceda_tuples["Target"] == "QCA"]
```

```
quakers_qca
```


	Source	Target	religion_name	first_year	last_year
21	Thomas (1) Hodgkin	QCA	Quaker	1839	1847
287	James Bowden	QCA	Quaker	1842	1847
288	William Nash	QCA	Quaker	1842	1847
289	Joseph Sturge	QCA	Quaker	1842	1847
290	William Jun Grimshaw	QCA	Quaker	1840	1847
291	Henry Knight	QCA	Quaker	1840	1847
293	Edward Paull	QCA	Quaker	1840	1847
294	Robert Jun Alsop	QCA	Quaker	1837	1847
295	Abram Rawlinson Barclay	QCA	Quaker	1837	1839
296	John Barclay	QCA	Quaker	1837	1839
297	Richard Barrett	QCA	Quaker	1837	1839
298	John Thomas Barry	QCA	Quaker	1837	1847
300	Peter Bedford	QCA	Quaker	1837	1847
302	John Bell	QCA	Quaker	1837	1839
304	Thomas Christy	QCA	Quaker	1837	1839
306	Samuel Darton	QCA	Quaker	1837	1839
307	Josiah Forster	QCA	Quaker	1837	1847
309	Robert Forster	QCA	Quaker	1837	1847
311	William Forster	QCA	Quaker	1837	1847
313	Joseph Talwin Foster	QCA	Quaker	1837	1847
314	John Hamilton	QCA	Quaker	1837	1839
315	Edwd Harris	QCA	Quaker	1837	1847
316	Geo Holmes	QCA	Quaker	1837	1839
317	Robert Howard	QCA	Quaker	1837	1839
319	John Kitching	QCA	Quaker	1837	1839
321	Joseph Neatby	QCA	Quaker	1837	1847
323	John Sanderson	QCA	Quaker	1837	1847
325	Joseph Shewell	QCA	Quaker	1837	1839
327	George Stacey	QCA	Quaker	1837	1847

	Source	Target	religion_name	first_year	last_year
328	Joseph Storrs	QCA	Quaker	1837	1847

```
quakers_qca.to_csv ('vw_hddt_ceda_quaker_qca.csv')
```

3.30 Quakers in the APS

```
quakers_aps = quakers_ceda_tuples[quakers_ceda_tuples["Target"] == "APS"]
```

```
quakers_aps.iloc [0:10]
```

	Source	Target	religion_name	first_year	last_year
0	William Spicer Wood	APS	Quaker	1864	1867
3	William Wilson	APS	Quaker	1838	1865
5	James Wilson	APS	Quaker	1862	1867
7	E T Wakefield	APS	Quaker	1853	1864
9	John Ross	APS	Quaker	1839	1852
10	J Robinson	APS	Quaker	1839	1840
12	William Horton Lloyd	APS	Quaker	1862	1862
14	Joseph Lister	APS	Quaker	1851	1855
16	Jonathan Hutchinson	APS	Quaker	1857	1866
19	William Holmes	APS	Quaker	1840	1867

```
quakers_aps.to_csv ('vw_hddt_ceda_quaker_aps.aps')
```

3.31 Quakers in the ESL

```
quakers_esl = quakers_ceda_tuples[quakers_ceda_tuples["Target"] == "ESL"]
```

```
quakers_esl
```

	Source	Target	religion_name	first_year	last_year
13	William Horton Lloyd	ESL	Quaker	1844	1847
15	Joseph Lister	ESL	Quaker	1844	1847
23	Thomas (1) Hodgkin	ESL	Quaker	1844	1862
25	John Henry Gurney	ESL	Quaker	1860	1867
29	Charles Henry Fox	ESL	Quaker	1861	1871
32	William Fowler	ESL	Quaker	1851	1851
34	Robert Nicholas Fowler	ESL	Quaker	1851	1871
39	David Dale	ESL	Quaker	1860	1863
44	x Collier	ESL	Quaker	1844	1844
46	William Clay	ESL	Quaker	1861	1868
48	Henry Christy	ESL	Quaker	1854	1865
58	James Bell	ESL	Quaker	1852	1862
60	James (1) Backhouse	ESL	Quaker	1869	1869
62	Edward Backhouse	ESL	Quaker	1870	1871
67	William Aldam	ESL	Quaker	1844	1848

```
quakers_esl.to_csv ('vw_hddt_ceda_quaker_esl.csv')
```

3.32 Quakers in the ASL

```
quakers_asl = quakers_ceda_tuples[quakers_ceda_tuples["Target"] == "ASL"]
```

```
quakers_asl
```

	Source	Target	religion_name	first_year	last_year
1	William Spicer Wood	ASL	Quaker	1863	1871
4	William Wilson	ASL	Quaker	1865	1866
6	James Wilson	ASL	Quaker	1865	1865
8	E T Wakefield	ASL	Quaker	1865	1868
11	J Robinson	ASL	Quaker	1865	1865
17	Jonathan Hutchinson	ASL	Quaker	1863	1871
20	William Holmes	ASL	Quaker	1865	1869
27	George Stacey Gibson	ASL	Quaker	1864	1866
37	James T J Doyle	ASL	Quaker	1865	1868
41	Henry Crowley	ASL	Quaker	1864	1871
50	Charles Buxton	ASL	Quaker	1864	1866
52	William Bull	ASL	Quaker	1867	1871
55	Antonio Brady	ASL	Quaker	1864	1871
65	S Stafford Allen	ASL	Quaker	1863	1870

```
quakers_asl.to_csv ('vw_hddt_ceda_quaker_asl.csv')
```

3.33 Quakers in the AI

```
quakers_ai = quakers_ceda_tuples[quakers_ceda_tuples["Target"] == "AI"]
```

```
quakers_ai
```

	Source	Target	religion_name	first_year	last_year
2	William Spicer Wood	AI	Quaker	1863	1871
18	Jonathan Hutchinson	AI	Quaker	1863	1871
30	Charles Henry Fox	AI	Quaker	1861	1871
35	Robert Nicholas Fowler	AI	Quaker	1851	1871
42	Henry Crowley	AI	Quaker	1864	1871
53	William Bull	AI	Quaker	1867	1871
56	Antonio Brady	AI	Quaker	1864	1871
63	Edward Backhouse	AI	Quaker	1870	1871

```
quakers_ai.to_csv ('vw_hddt_ceda_quaker_ai.csv')
```

P7 Chapter 4 Case Study 1 The Centres for the Emergence of the Discipline of Anthropology (CEDA)

File name: jnb_hddt_ceda_dyn_edges

4.1 HDDT Visualisations - CEDA bigraph

This project explores 4 of 5 'foundation societies' recognised by RAI, and 1 'origin' society (the CQA)

add by me.

Society	abv.	Dates
Quaker Committee on the Aborigines*	QCA	1832/37 - 1846
Aborigines Protection Society	APS	1837 - 1919
Ethnological Society of London	ESL	1843 - 1871
Anthropological Society of London	ASL	1863 - 1871
Anthropological Institute	AI	1843 - 1871
London Anthropological Society**	LAS	1873 - 1874

- Origin Society included in this project but not recognised by RAI. ** not included in this project (beyond 1871 cut off date).

```
import csv
from operator import itemgetter
import networkx as nx
from networkx.algorithms import community #This part of networkx, for commun
#needs to be imported separately.
import nbconvert
import pandas as pd
import matplotlib.pyplot as plt
plt.rcParams["figure.figsize"] = (12, 6)
import seaborn as sn
import numpy as np

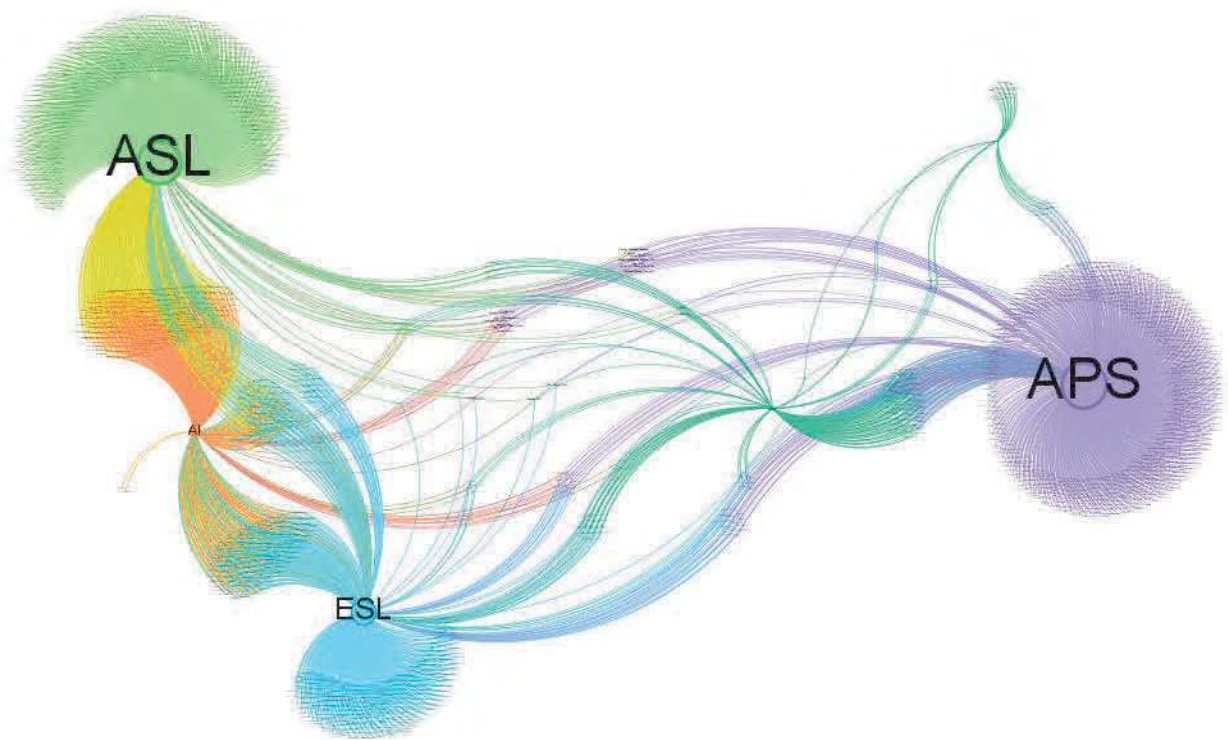
plt.rcParams.update({'font.size': 18})

plt.rc('figure', figsize=(20, 10))
```

```
-----
ModuleNotFoundError                                Traceback (most recent call last)
Cell In[1], line 3
      1 import csv
      2 from operator import itemgetter
----> 3 import networkx as nx
      4 from networkx.algorithms import community #This part of networkx, fo
      5 #needs to be imported separately.

ModuleNotFoundError: No module named 'networkx'
```

4.2 The CEDA 1830 - 1870



4.3 The Quaker Committee on the Aborigines (QCA) 1837 -1846

Because there are only 30 members of the QCA we can list all of them here

The Quaker Committee on the Aborigines, was a Quaker Committee of Enquiry. A committee formed by, and exclusively manned by Quakers. It met, performed its enquiries and reported its findings and recommendations to the Quaker Meetings for Sufferings, the

standing committee of London Yearly Meeting which was the National Assembly of Quakers in Britain at the time. The committees remit, rules of engagement and characteristics would have been agreed by the national assembly and the committee would no doubt have reported in the manner of Friends. The committee was formed to explore and take up a 'concern' amongst Quakers, initially to consider promoting the Gospel amongst the aborigines in 1832 (prompted by similar actions popular at the time among other evangelical churches). But it changed its remit in 1837 to instead take up a philanthropic concern deriving from the group's increasing awareness through its activities of the plight of aborigines. Therefore, what began as a Quaker Committee of Enquiry to consider promoting the Gospel to Aborigines, soon transformed into the Quaker Committee on the Aborigines, concerned with the plight of the aborigines throughout the colonies, and it's relief.

```
qca = pd.read_csv ('vw_4_ceda_membership_quakers_qca2.csv')
qca['birth_year'] = qca ['birth_year'].fillna(0).astype(np.int64)
qca['death_year'] = qca['death_year'].fillna(0).astype(np.int64)
qca.info ()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30 entries, 0 to 29
Data columns (total 7 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Name                                  30 non-null     object
1   birth_year                           30 non-null     int64
2   death_year                           30 non-null     int64
3   religion_name                         30 non-null     object
4   ceda_name                             30 non-null     object
5   person_ceda_first_year                30 non-null     int64
6   person_ceda_last_year                 30 non-null     int64
dtypes: int64(4), object(3)
memory usage: 1.8+ KB
```

```
qca
```

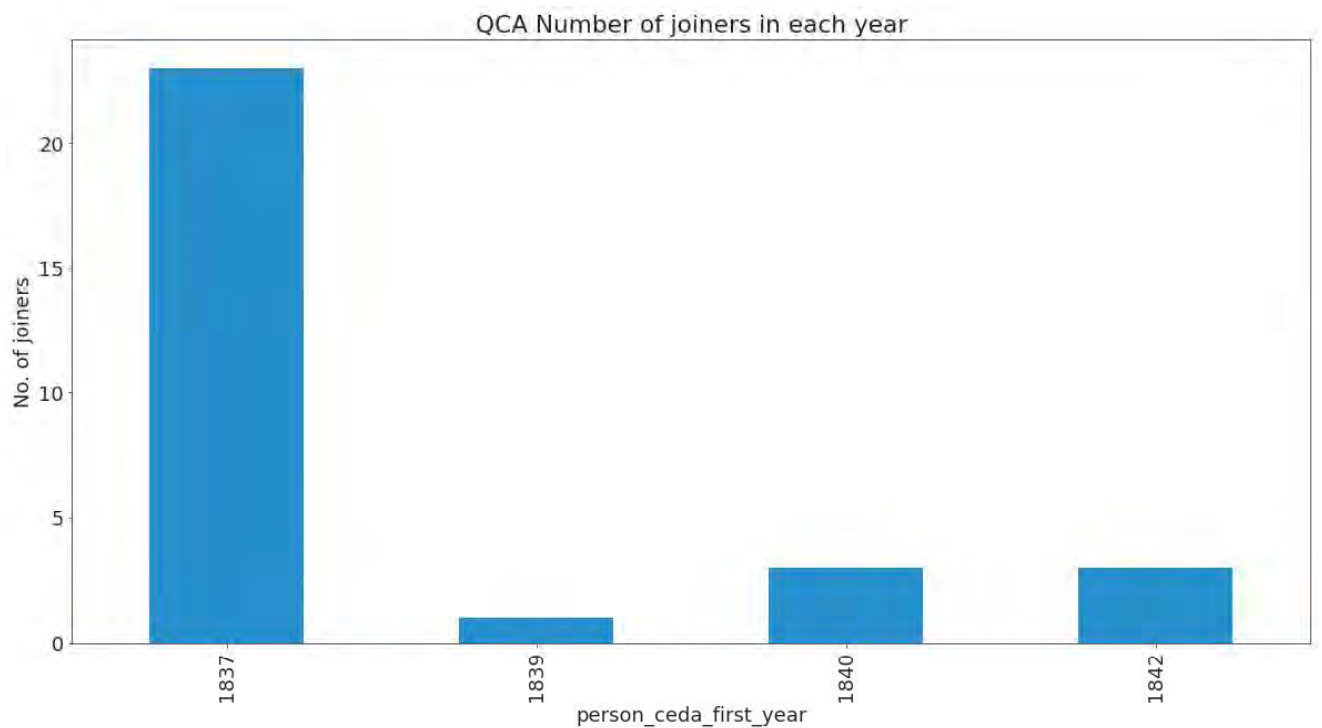

	Name	birth_year	death_year	religion_name	ceda_name	person_co
0	Thomas (1) Hodgkin	1798	1866	Quaker	QCA	
1	James Bowden	0	0	Quaker	QCA	
2	William Nash	0	0	Quaker	QCA	
3	Joseph Sturge	0	0	Quaker	QCA	
4	William Jun Grimshaw	0	0	Quaker	QCA	
5	Henry Knight	0	0	Quaker	QCA	
6	Edward Paull	0	0	Quaker	QCA	
7	Robert Jun Alsop	0	0	Quaker	QCA	
8	Abram Rawlinson Barclay	0	0	Quaker	QCA	
9	John Barclay	0	0	Quaker	QCA	
10	Richard Barrett	0	0	Quaker	QCA	
11	John Thomas Barry	0	0	Quaker	QCA	
12	Peter Bedford	0	0	Quaker	QCA	
13	John Bell	0	0	Quaker	QCA	
14	Thomas Christy	0	0	Quaker	QCA	
15	Samuel Darton	0	0	Quaker	QCA	
16	Josiah Forster	0	0	Quaker	QCA	
17	Robert Forster	0	0	Quaker	QCA	

	Name	birth_year	death_year	religion_name	ceda_name	person_co
18	William Forster	0	0	Quaker	QCA	
19	Joseph Talwin Foster	0	0	Quaker	QCA	
20	John Hamilton	0	0	Quaker	QCA	
21	Edwd Harris	0	0	Quaker	QCA	
22	Geo Holmes	0	0	Quaker	QCA	
23	Robert Howard	0	0	Quaker	QCA	
24	John Kitching	0	0	Quaker	QCA	
25	Joseph Neatby	0	0	Quaker	QCA	
26	John Sanderson	0	0	Quaker	QCA	
27	Joseph Shewell	0	0	Quaker	QCA	
28	George Stacey	0	0	Quaker	QCA	
29	Joseph Storrs	0	0	Quaker	QCA	

4.4 CQA Joiners each year

We can plot the number of joiners in each year. New members joined only in the years 1837, 1839, 1840 and 1842. We know that the QCA was established in 1837, so there were only three years when new members joined.

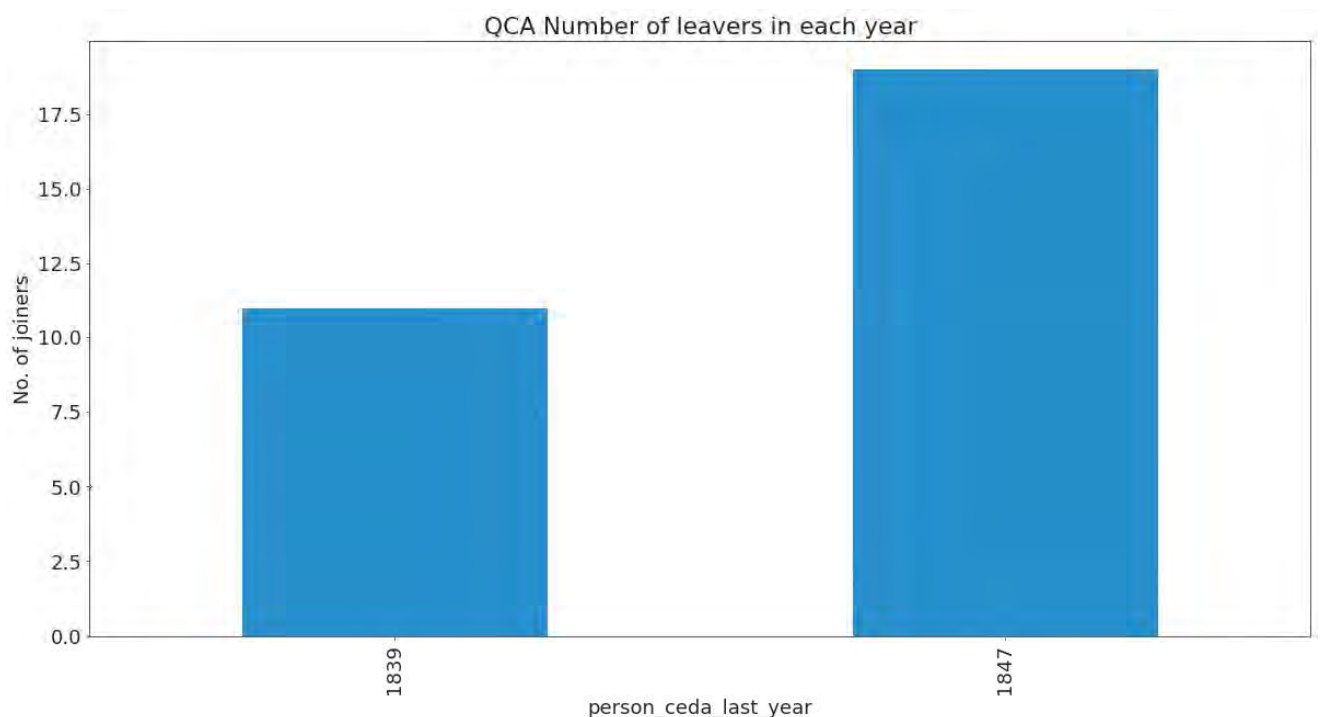
```
qca.groupby('person_ceda_first_year')['Name'].nunique().plot(kind='bar')
plt.title("QCA Number of joiners in each year")
plt.ylabel("No. of joiners")
plt.show()
```



4.5 CQA Leavers each year

Because new members group in specific years, we chart the number of leavers in each year. Members left in 1839, 1842 and 1847. we know that the QCA was 'laid down' (disbanded) in 1847 which leaves only two years when members left the committee.

```
qca.groupby('person_ceda_last_year')['Name'].unique().plot(kind='bar')
plt.title("QCA Number of leavers in each year")
plt.ylabel("No. of joiners")
plt.show()
```



We can elsewhere investigate why in 1839, 11 members left and 1 joined (Thomas Hodgkin). 3 more joined in each of 1840 and 1842.

4.6 Duration in the CEDA

```
qca[(qca['person_ceda_first_year'] == 1837) & (qca['person_ceda_last_year']
```

	Name	birth_year	death_year	religion_name	ceda_name	person_co
7	Robert Jun Alsop	0	0	Quaker	QCA	
11	John Thomas Barry	0	0	Quaker	QCA	
12	Peter Bedford	0	0	Quaker	QCA	
16	Josiah Forster	0	0	Quaker	QCA	
17	Robert Forster	0	0	Quaker	QCA	
18	William Forster	0	0	Quaker	QCA	
19	Joseph Talwin Foster	0	0	Quaker	QCA	
21	Edwd Harris	0	0	Quaker	QCA	
25	Joseph Neatby	0	0	Quaker	QCA	
26	John Sanderson	0	0	Quaker	QCA	
28	George Stacey	0	0	Quaker	QCA	
29	Joseph Storrs	0	0	Quaker	QCA	

12 of the original members were members throughout the life of the committee. 11 of the original members left after the first year. 3 new members in 1840 and 3 new members in [1841](#). In any year the majority of members were 'permanent' members.

4.7 The Aborigines Protection Society (APS) 1837 -1919

The database contains the names of 1171 members of the APS from its foundation in 1838 to 1871 when it merged with Anti-Slavery International. 571 members (49%) are Quaker.

The Aborigines Protection Society was a secular pressure group that lobbied the Colonial Office and Parliament for the relief of the plight of aborigines throughout the British Settlements. It had a mixed Quaker and non-Quaker executive, membership and subscription lists (it was in large part drawn from the Quaker Committee on the Aborigines), and Quakers dominated the agenda and publishing and lobbying activities of the society for at least the first 30 years of the Society's life. The Society met, performed its enquiries and reported its findings and recommendations to the Society's members according to its own constitution (it usually met monthly). The Society's remit, rules of engagement and characteristics were similar to those of the many other secular lobbying and public opinion forming societies of the time.

```
aps = pd.read_csv ('vw_4_ceda_membership_dates_aps.csv')
aps['birth_year'] = aps ['birth_year'].fillna(0).astype(np.int64)
aps['death_year'] = aps['death_year'].fillna(0).astype(np.int64)
aps.info ()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1170 entries, 0 to 1169
Data columns (total 6 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Name             1170 non-null   object
1   birth_year       1170 non-null   int64
2   death_year       1170 non-null   int64
3   Target           1170 non-null   object
4   first_year       1170 non-null   int64
5   last_year        1170 non-null   int64
dtypes: int64(4), object(2)
memory usage: 55.0+ KB
```

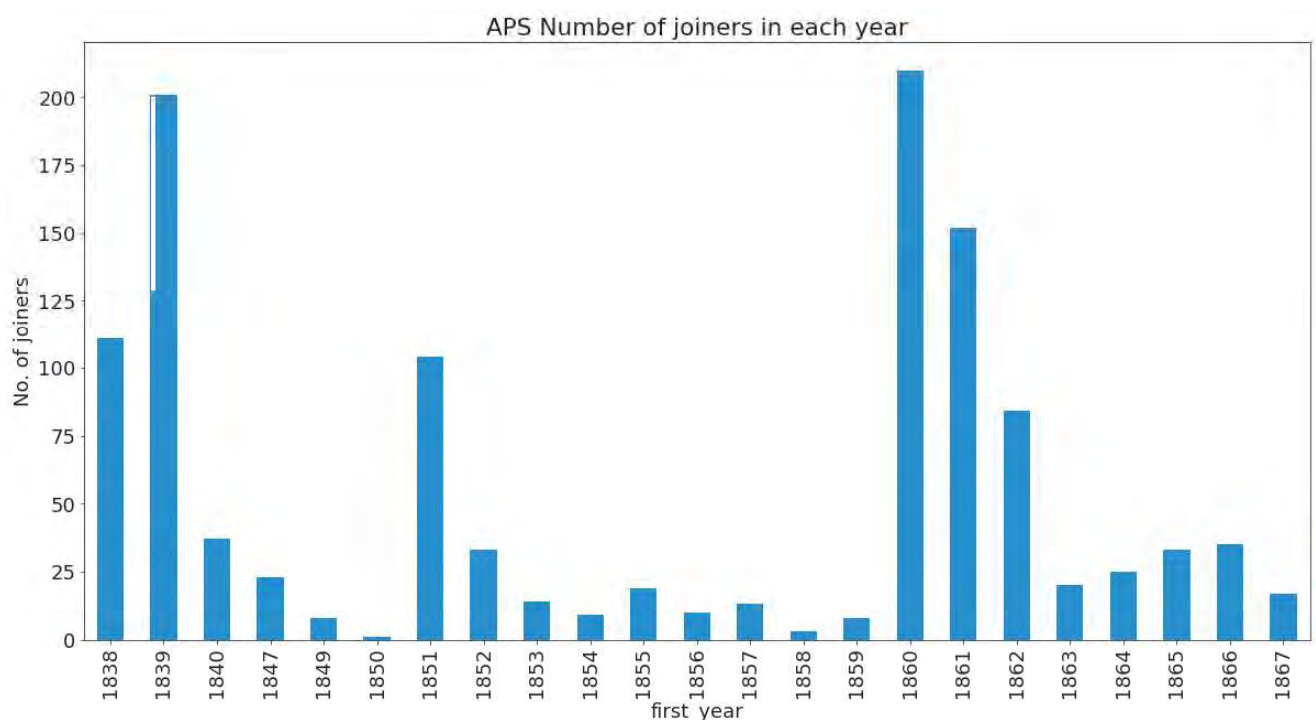
```
aps
```

	Name	birth_year	death_year	Target	first_year	last_year
0	William Aldam	1813	1890	APS	1838	1848
1	Samuel C Baker	1821	1893	APS	1839	1871
2	James Bell	1818	1872	APS	1847	1862
3	John Bell (2)	1811	1895	APS	1838	1855
4	John Brown	1801	1879	APS	1839	1839
...
1165	x Wright	0	0	APS	1839	1850
1166	W Wrigley	0	0	APS	1861	1862
1167	James Yates	0	0	APS	1853	1856
1168	John Young	0	0	APS	1840	1840
1169	Thomas Zachary	0	0	APS	1840	1867

1170 rows × 6 columns

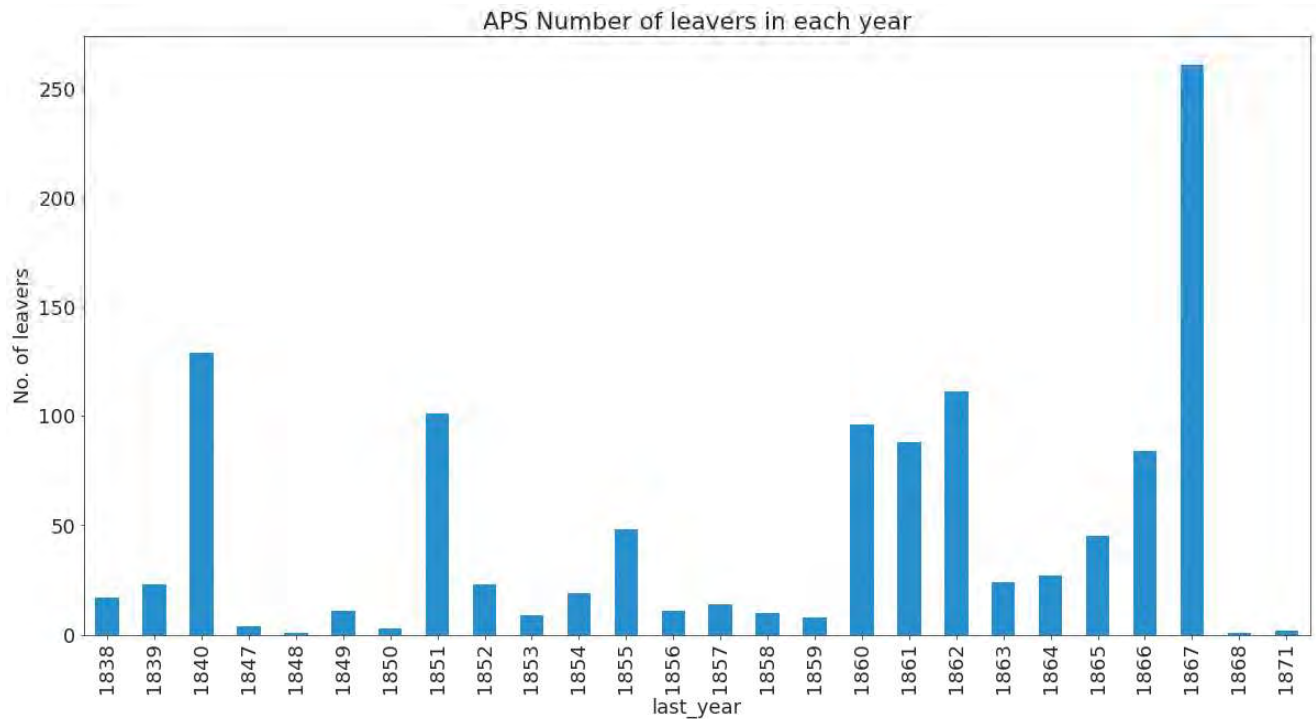
4.8 APS joiners in each year

```
aps.groupby('first_year')['Name'].nunique().plot(kind='bar')
plt.title("APS Number of joiners in each year")
plt.ylabel("No. of joiners")
plt.show()
```



4.9 APS leavers in each year

```
aps.groupby('last_year')['Name'].nunique().plot(kind='bar')
plt.title("APS Number of leavers in each year")
plt.ylabel("No. of leavers")
plt.show()
```



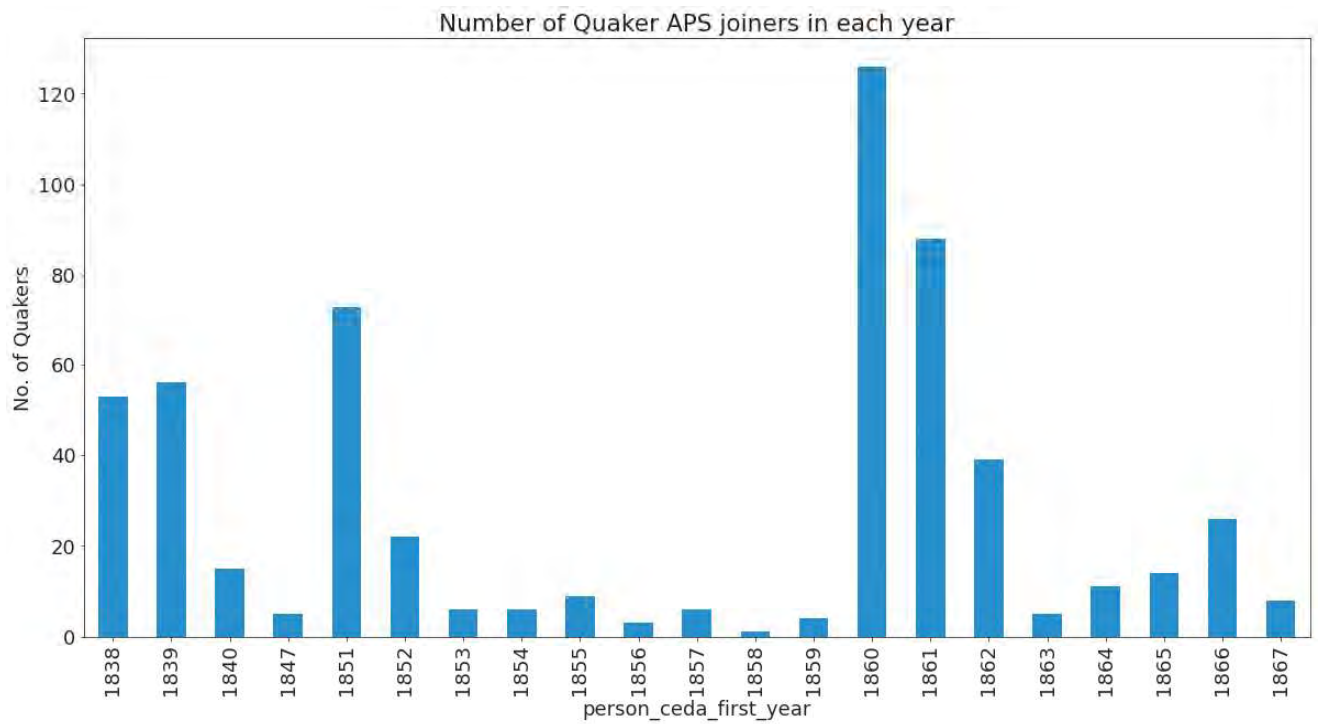
```
quakers_aps = pd.read_csv ('vw_4_ceda_membership_quakers_aps2.csv')
# quakers_aps = aps.loc[aps ['religion_name'] == 'Quaker',:]
quakers_aps
```

	Name	birth_year	death_year	religion_name	ceda_name	person_u
0	William Spicer Wood	NaN	1902.0	Quaker	APS	
1	William Wilson	1785.0	1868.0	Quaker	APS	
2	James Wilson	NaN	NaN	Quaker	APS	
3	E T Wakefield	NaN	NaN	Quaker	APS	
4	John Ross	NaN	NaN	Quaker	APS	
...	
571	Joshua Wilson	NaN	NaN	Quaker	APS	
572	F Woodhead	NaN	NaN	Quaker	APS	
573	W Woolston	NaN	NaN	Quaker	APS	
574	Francis Wright	NaN	NaN	Quaker	APS	
575	S W Wright	NaN	NaN	Quaker	APS	

576 rows x 7 columns

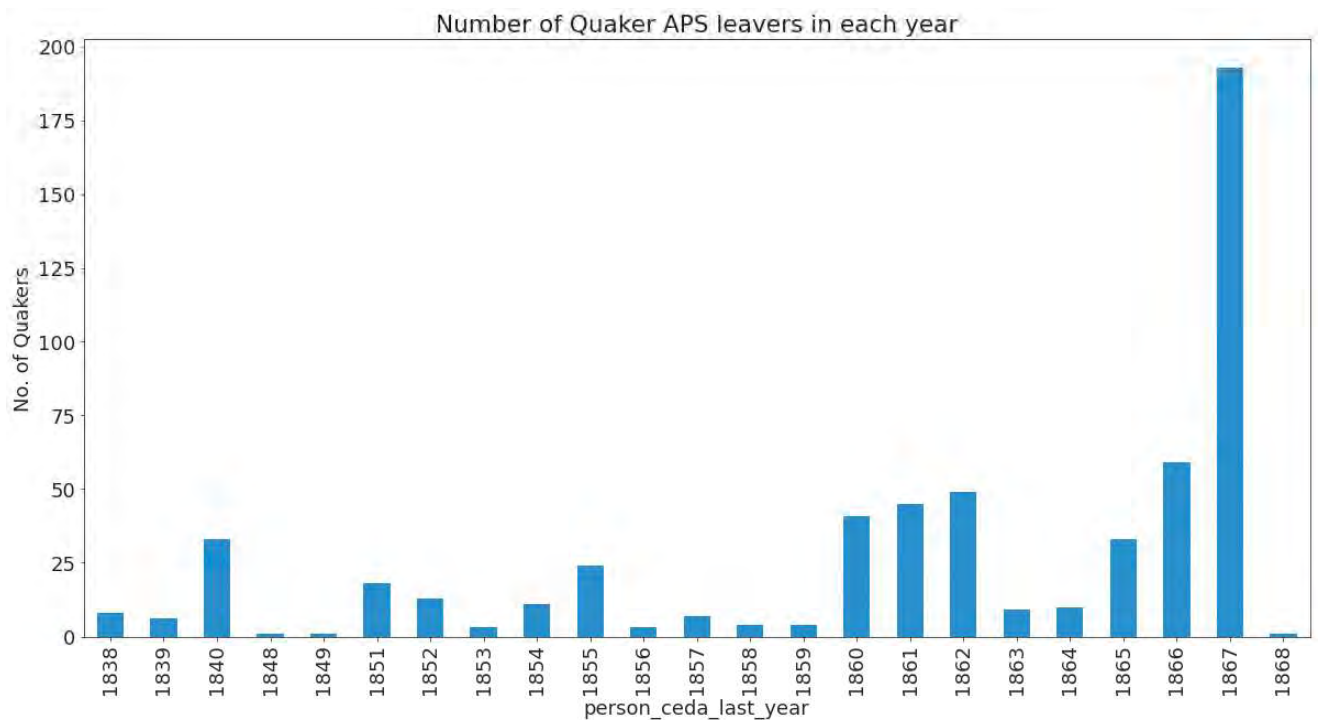
4.10 Quakers joining the APS in each year

```
quakers_aps.groupby('person_ceda_first_year')['Name'].nunique().plot(kind='line')
plt.title ("Number of Quaker APS joiners in each year")
plt.ylabel ("No. of Quakers")
plt.show()
```

4.11 Quakers leaving the APS in each year

```
quakers_aps.groupby('person_ceda_last_year')['Name'].nunique().plot(kind='bar')
plt.title("Number of Quaker APS leavers in each year")
plt.ylabel("No. of Quakers")
plt.show()
```



4.12 The Ethnological Society of London (ESL) 1843 - 1871

The Ethnological Society of London was the first intentionally academic society devoted to the discipline of anthropology in Britain. Secular by intent but if not always entirely so in its early years, it sought to be a place where those with a scientific interest in the field of ethnology could commune, share ideas and knowledge, and produce academic reports and hold academic meetings. It met, performed its enquiries and reported its findings and recommendations to the Society's members according to its own constitution (it usually met monthly). The Society's remit, rules of engagement and characteristics were those of the many other scientific societies emerging at the time, its constitution being purposely compliant with BAAS requirements.

```
esl = pd.read_csv ('vw_4_ceda_membership_dates_esl.csv')

# code not needed for this set because in this dataframe birth_year and death_year are not null
#esl['birth_year'] = esl['birth_year'].fillna(0).astype(np.int64)
#esl['death_year'] = esl['death_year'].fillna(0).astype(np.int64)
# esl['religion_name'] = esl['religion_name'].fillna(0).astype(np.int64)
esl.info ()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 748 entries, 0 to 747
Data columns (total 6 columns):
#   Column      Non-Null Count  Dtype
---  -
0   Name         748 non-null    object
1   birth_year   436 non-null    object
2   death_year   466 non-null    object
3   Target       748 non-null    object
4   first_year   748 non-null    int64
5   last_year    748 non-null    int64
dtypes: int64(2), object(4)
memory usage: 35.2+ KB
```

```
esl
```

	Name	birth_year	death_year	Target	first_year	last_year
0	William Adam	NaN	NaN	ESL	1844	1844
1	William (1) Adams	NaN	NaN	ESL	1844	1844
2	William (2) Adams	1,820	1,900	ESL	1858	1871
3	Louis Agassiz	1,807	1,873	ESL	1860	1871
4	Augustine Aglio	1,777	1,857	ESL	1843	1845
...
743	James Wyld	1,812	1,887	ESL	1844	1854
744	Ashton Yates	1,781	1,863	ESL	1860	1862
745	W Holt Yates	1,802	1,874	ESL	1844	1846
746	James Yearsley	1,805	1,869	ESL	1845	1845
747	Arthur de Zeltner	NaN	NaN	ESL	1865	1871

748 rows x 6 columns

```
quakers_esl = pd.read_csv ('vw_4_ceda_membership_quakers_esl2.csv')
quakers_esl
```

	Name	birth_year	death_year	religion_name	ceda_name	person_co
0	William Horton Lloyd	NaN	NaN	Quaker	ESL	
1	Joseph Lister	1827.0	1912.0	Quaker	ESL	
2	Thomas (1) Hodgkin	1798.0	1866.0	Quaker	ESL	
3	John Henry Gurney	1819.0	1890.0	Quaker	ESL	
4	Charles Henry Fox	NaN	NaN	Quaker	ESL	
5	William Fowler	NaN	NaN	Quaker	ESL	
6	Robert Nicholas Fowler	1828.0	1891.0	Quaker	ESL	
7	David Dale	1829.0	1906.0	Quaker	ESL	
8	x Collier	NaN	NaN	Quaker	ESL	
9	William Clay	1791.0	1869.0	Quaker	ESL	
10	Henry Christy	1810.0	1865.0	Quaker	ESL	
11	James Bell	1818.0	1872.0	Quaker	ESL	
12	James (1) Backhouse	1794.0	1869.0	Quaker	ESL	
13	Edward Backhouse	1808.0	1879.0	Quaker	ESL	
14	William Aldam	1813.0	1890.0	Quaker	ESL	

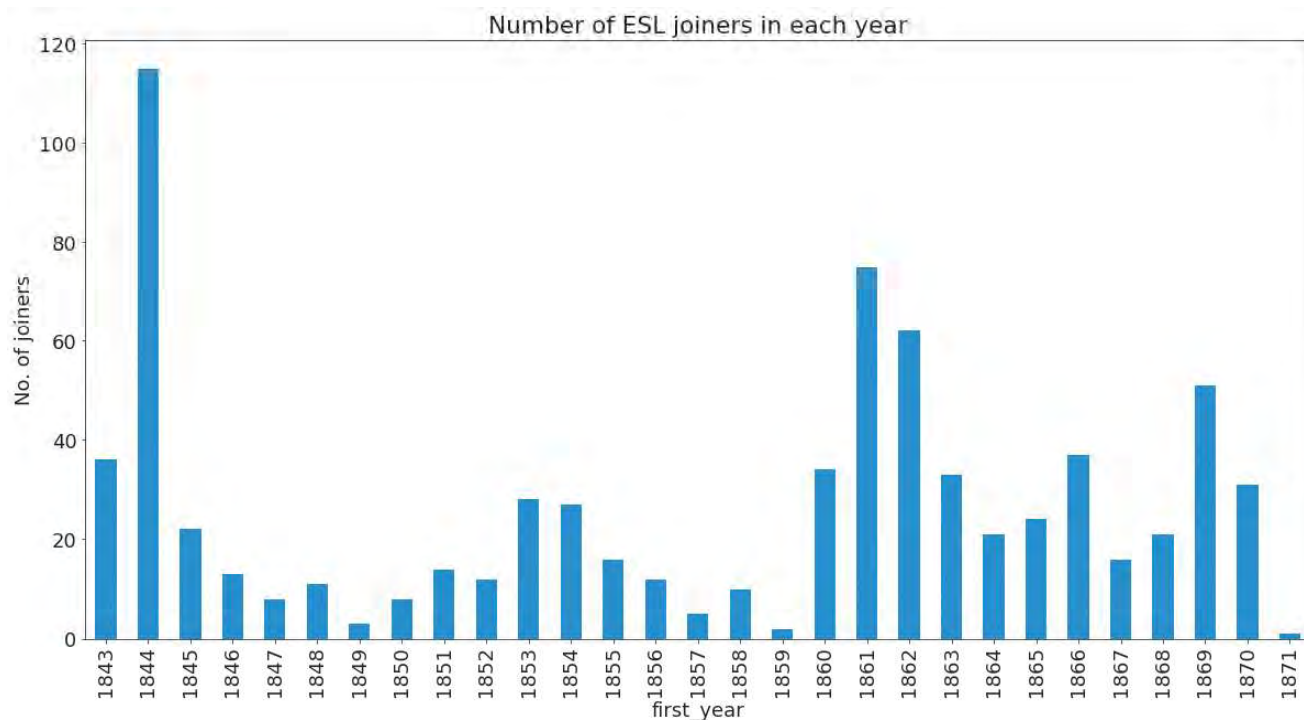
 image.png

4.13 ESL joiners in each year

```

esl.groupby('first_year')['Name'].nunique().plot(kind='bar')
plt.title ("Number of ESL joiners in each year")
plt.ylabel ("No. of joiners")
plt.show()

```

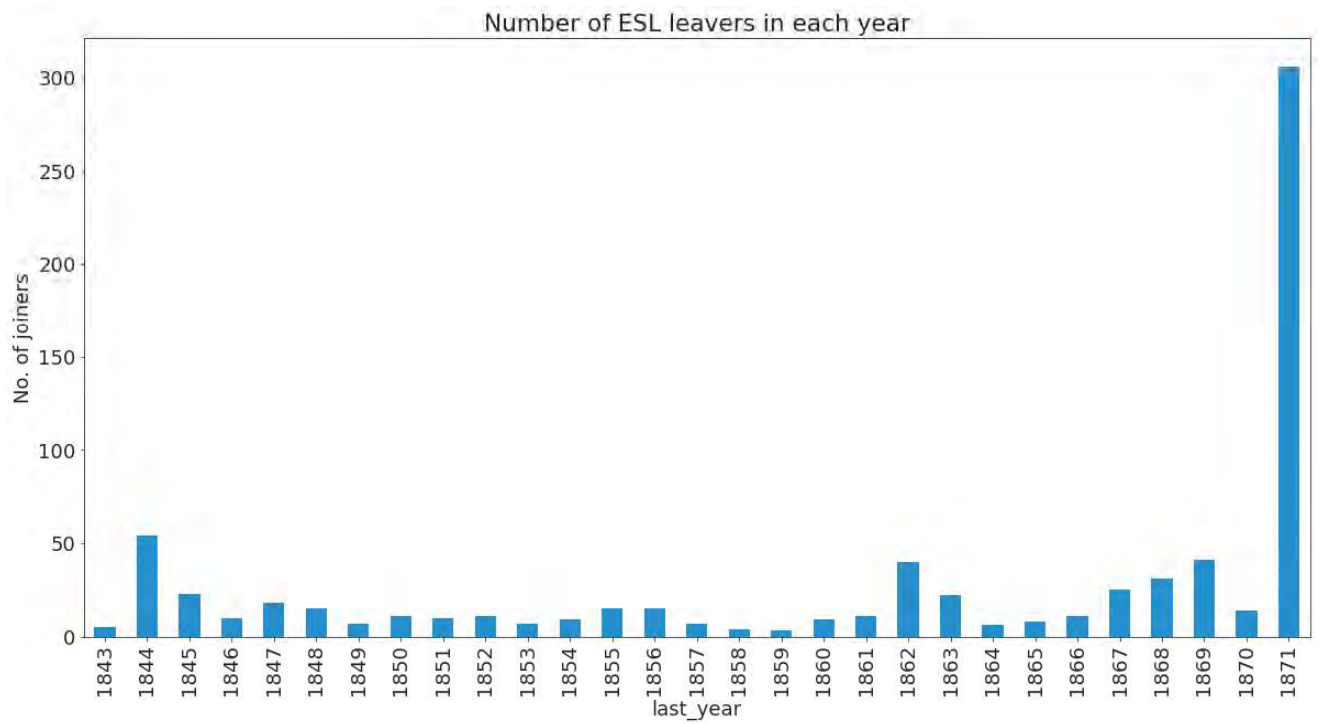


4.14 ESL leavers in each year

```

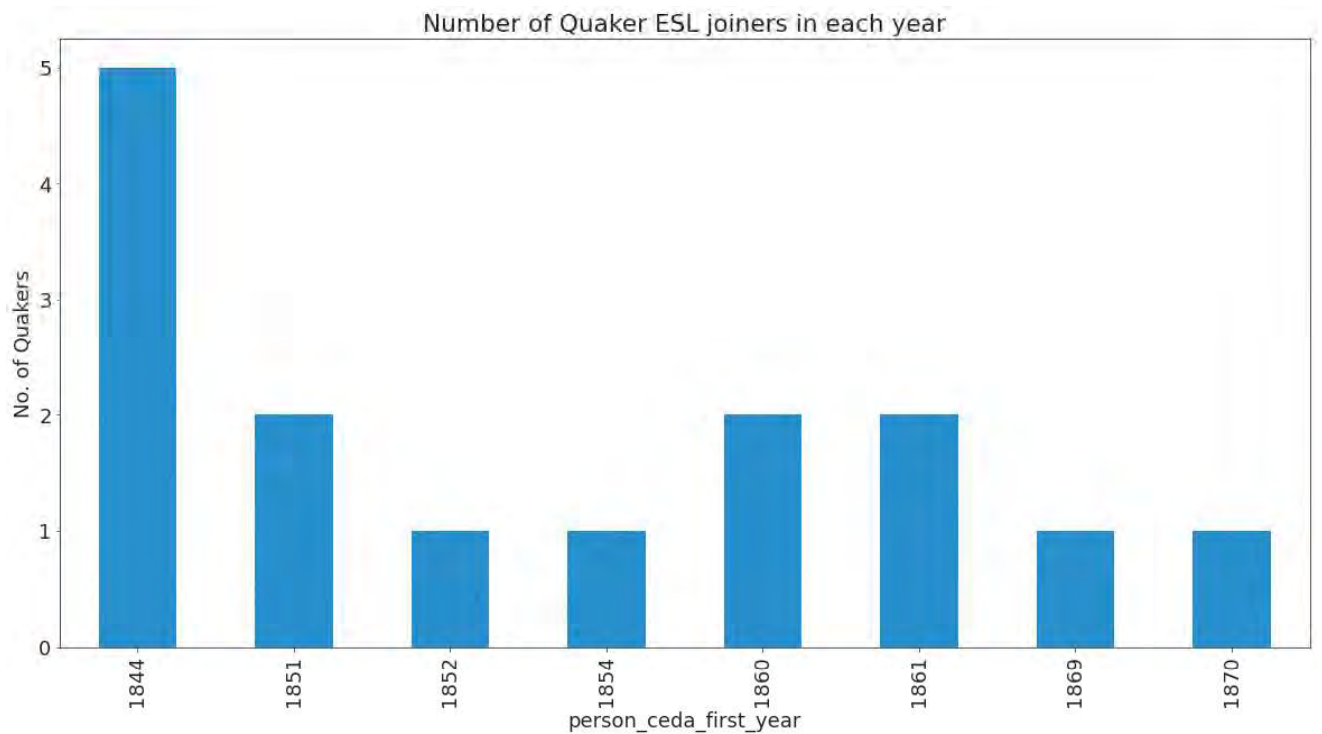
esl.groupby('last_year')['Name'].nunique().plot(kind='bar')
plt.title ("Number of ESL leavers in each year")
plt.ylabel ("No. of joiners")
plt.show()

```



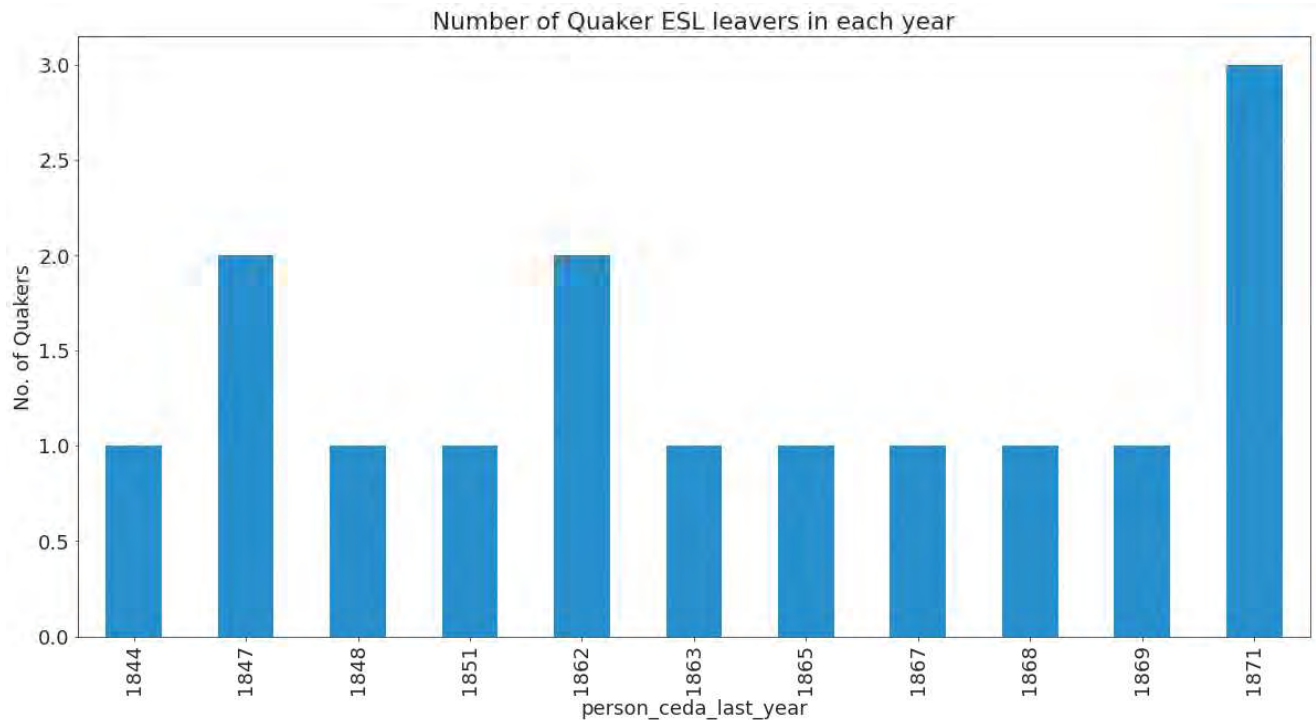
4.15 ESL Quaker joiners in each year

```
quakers_esl.groupby('person_ceda_first_year')['Name'].nunique().plot(kind='bar')
plt.title("Number of Quaker ESL joiners in each year")
plt.ylabel("No. of Quakers")
plt.show()
```



4.16 ESL Quaker leavers in each year

```
quakers_esl.groupby('person_ceda_last_year')['Name'].nunique().plot(kind='bar')
plt.title("Number of Quaker ESL leavers in each year")
plt.ylabel("No. of Quakers")
plt.show()
```



4.17 The Anthropological Society of London (ASL) 1863 - 1871

```
asl = pd.read_csv('vw_4_ceda_membership_dates_asl.csv')
asl['birth_year'] = asl['birth_year'].fillna(0).astype(np.int64)
asl['death_year'] = asl['death_year'].fillna(0).astype(np.int64)
```

```
asl.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1334 entries, 0 to 1333
Data columns (total 6 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Name             1334 non-null   object
1   birth_year       1334 non-null   int64
2   death_year       1334 non-null   int64
3   Target           1334 non-null   object
4   first_year       1334 non-null   int64
5   last_year        1334 non-null   int64
dtypes: int64(4), object(2)
memory usage: 62.7+ KB
```

```
asl
```

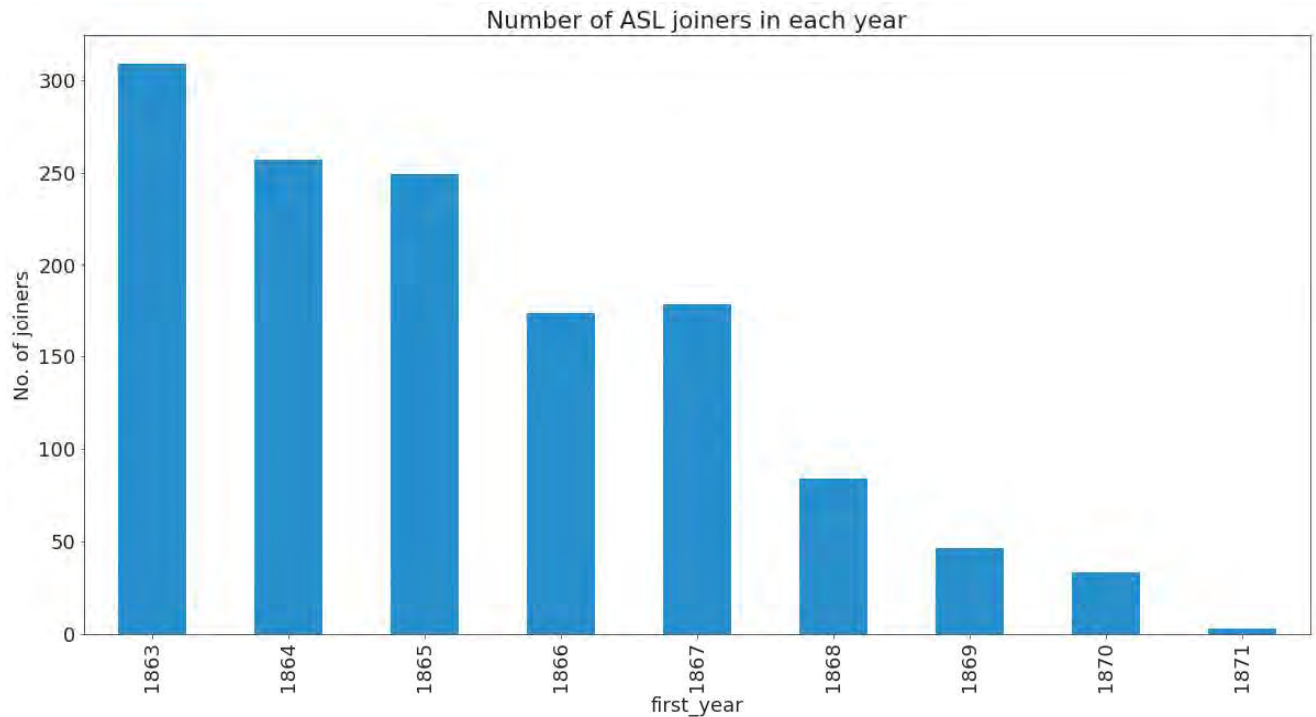
	Name	birth_year	death_year	Target	first_year	last_year
0	Arthur William A Beckett	1844	1909	ASL	1864	1867
1	Andrew Mercer Adam	0	0	ASL	1865	1867
2	H R Adam	0	0	ASL	1870	1871
3	Henry John Adams	0	0	ASL	1864	1869
4	William Adlam	0	0	ASL	1863	1866
...
1329	Stephen Yeldham	1810	1896	ASL	1866	1869
1330	James A Youl	1811	1904	ASL	1864	1865
1331	Robert Younge	1801	1874	ASL	1865	1871
1332	Arthur de Zeltner	0	0	ASL	1866	1871
1333	x Zohrab	0	0	ASL	1867	1871

1334 rows x 6 columns

```
quakers_asl = pd.read_csv ('vw_4_ceda_membership_quakers_asl2.csv')
```

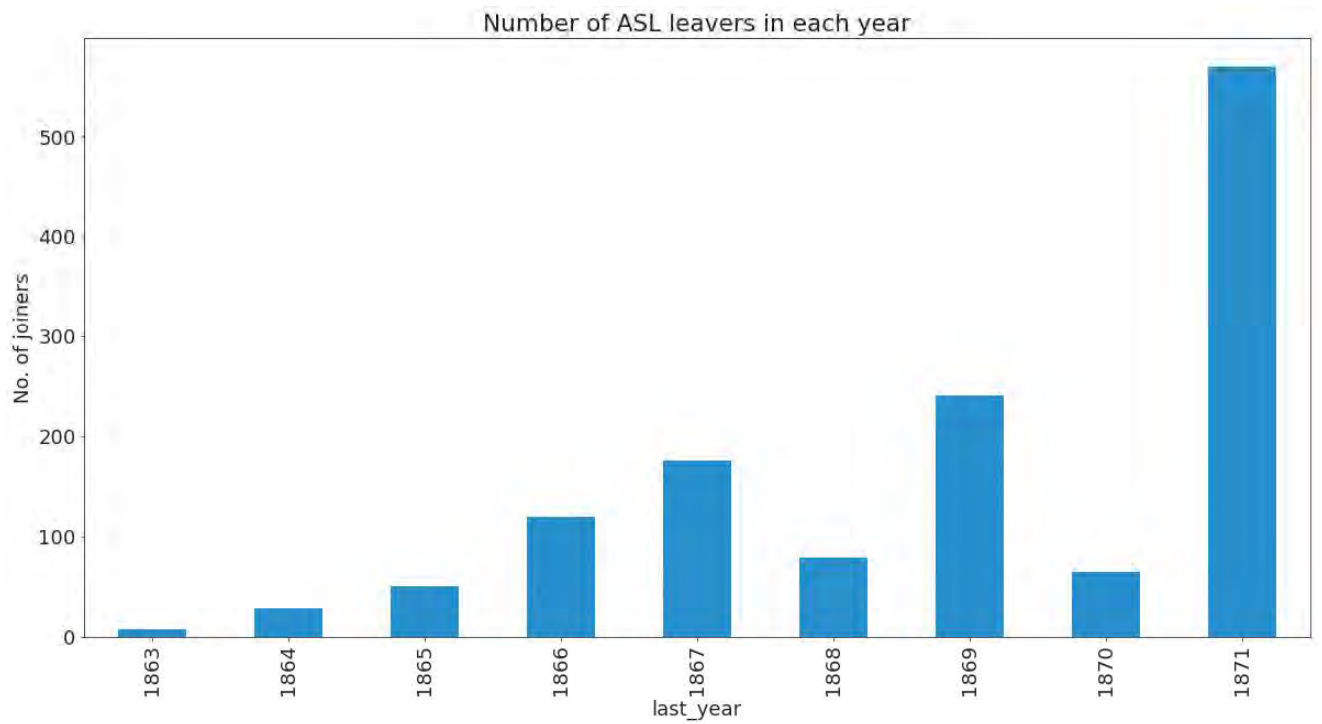
4.18 ASL joiners in each year


```
asl.groupby('first_year')['Name'].nunique().plot(kind='bar')
plt.title ("Number of ASL joiners in each year")
plt.ylabel ("No. of joiners")
plt.show()
```



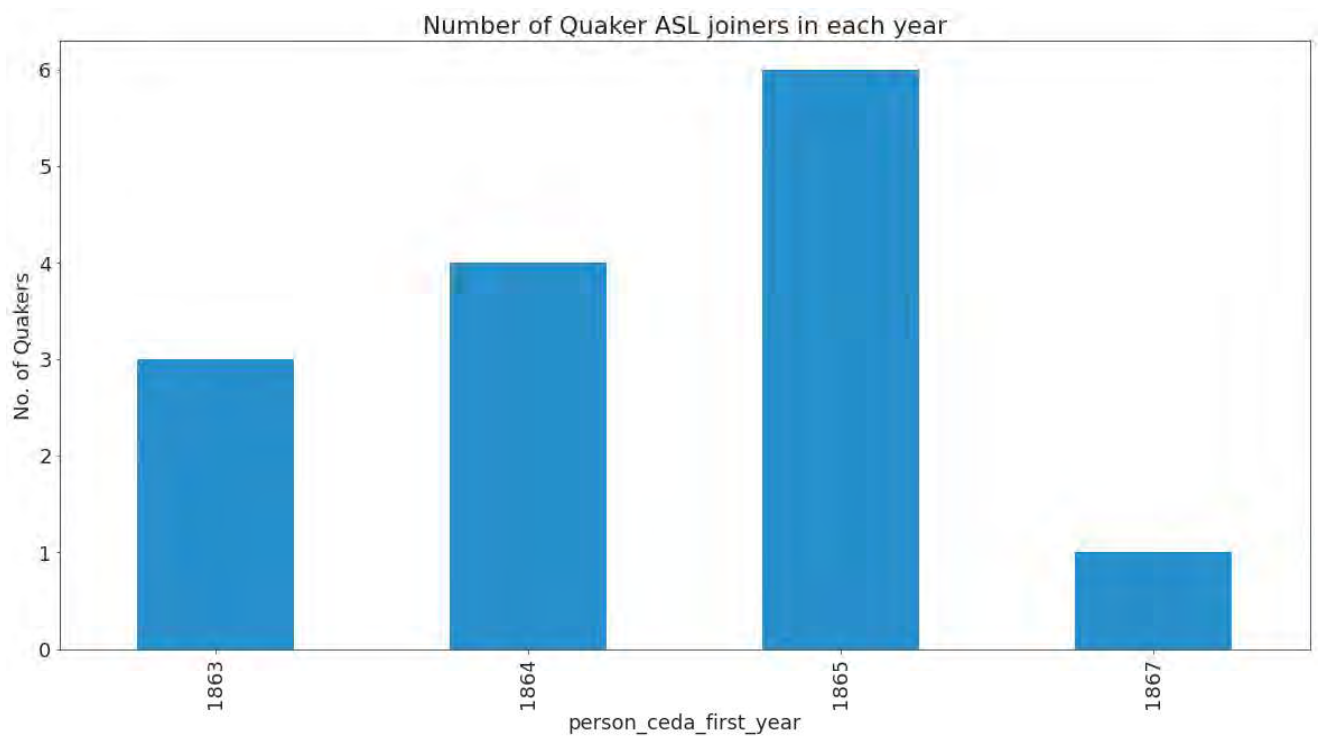
4.19 ASL leavers in each year

```
asl.groupby('last_year')['Name'].nunique().plot(kind='bar')
plt.title ("Number of ASL leavers in each year")
plt.ylabel ("No. of joiners")
plt.show()
```



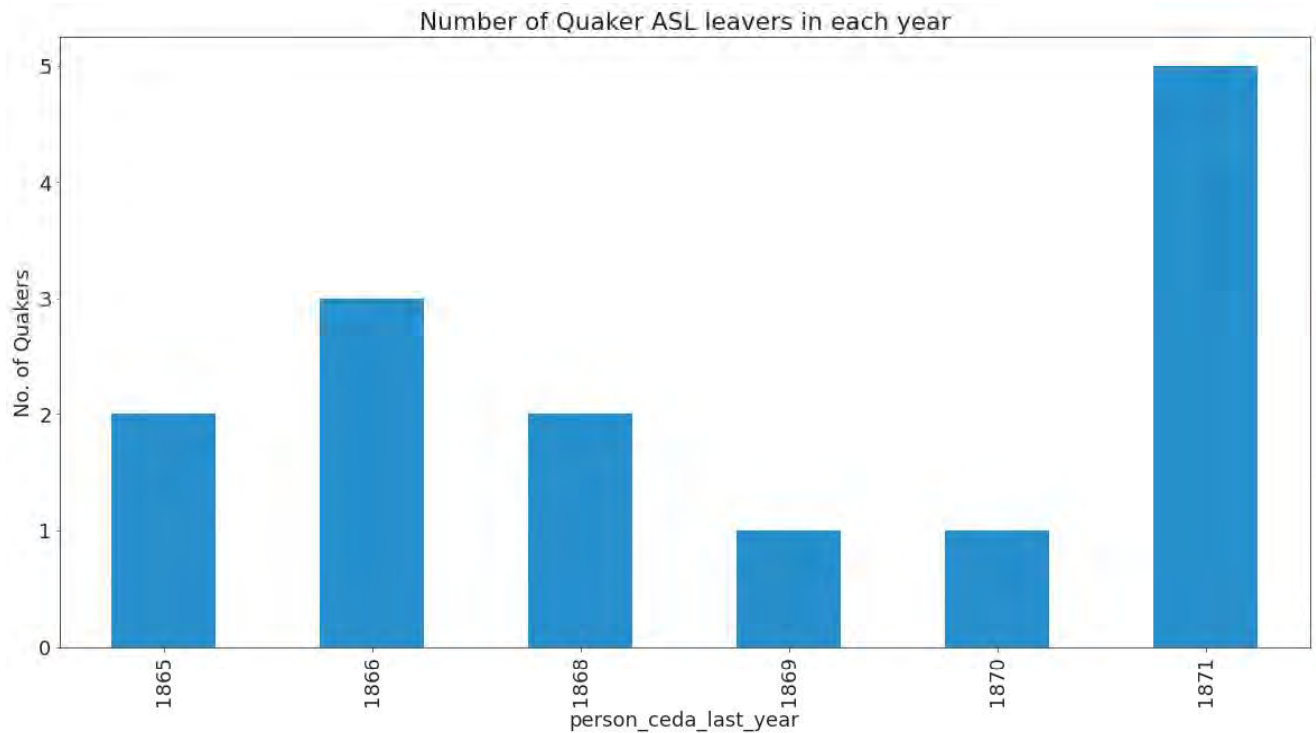
4.20 ASL Quaker joiners in each year

```
quakers_asl.groupby('person_ceda_first_year')['Name'].nunique().plot(kind='bar')  
plt.title("Number of Quaker ASL joiners in each year")  
plt.ylabel("No. of Quakers")  
plt.show()
```



4.21 ASL Quaker leavers in each year

```
quakers_asl.groupby('person_ceda_last_year')['Name'].nunique().plot(kind='bar')
plt.title("Number of Quaker ASL leavers in each year")
plt.ylabel("No. of Quakers")
plt.show()
```



```
quakers_asl = pd.read_csv ('vw_4_ceda_membership_quakers_asl2.csv')
quakers_asl
```

	Name	birth_year	death_year	religion_name	ceda_name	person_c
0	William Spicer Wood	NaN	1902.0	Quaker	ASL	
1	William Wilson	1785.0	1868.0	Quaker	ASL	
2	James Wilson	NaN	NaN	Quaker	ASL	
3	E T Wakefield	NaN	NaN	Quaker	ASL	
4	J Robinson	NaN	NaN	Quaker	ASL	
5	Jonathan Hutchinson	1828.0	1913.0	Quaker	ASL	
6	William Holmes	NaN	NaN	Quaker	ASL	
7	George Stacey Gibson	1818.0	1883.0	Quaker	ASL	
8	James T J Doyle	NaN	NaN	Quaker	ASL	
9	Henry Crowley	NaN	1887.0	Quaker	ASL	
10	Charles Buxton	1823.0	1871.0	Quaker	ASL	
11	William Bull	1828.0	1902.0	Quaker	ASL	
12	Antonio Brady	1811.0	1881.0	Quaker	ASL	
13	S Stafford Allen	1840.0	1870.0	Quaker	ASL	

 image.png

4.22 Anthropological Institute (AI) 1843 - 1871

```
ai = pd.read_csv ('vw_4_ceda_membership_dates_ai.csv')
# code not needed for this set because in this dataframe birth_year and death_year are not null
#ai['birth_year'] = ai ['birth_year'].fillna(0).astype(np.int64)
#ai['death_year'] = ai['death_year'].fillna(0).astype(np.int64)
```

```
ai.info ()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 610 entries, 0 to 609
Data columns (total 6 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Name             610 non-null    object
1   birth_year       399 non-null    object
2   death_year       436 non-null    object
3   Target           610 non-null    object
4   first_year       610 non-null    int64
5   last_year        610 non-null    int64
dtypes: int64(2), object(4)
memory usage: 28.7+ KB
```

```
ai
```

	Name	birth_year	death_year	Target	first_year	last_year
0	H R Adam	NaN	NaN	AI	1870	1871
1	William (2) Adams	1,820	1,900	AI	1858	1871
2	Louis Agassiz	1,807	1,873	AI	1860	1871
3	Alexander Muirhead Aitken	NaN	NaN	AI	1864	1871
4	William Amhurst Tyssen Amhurst	1,835	1,909	AI	1862	1871
...
605	Robert Carr Woods	1,816	1,875	AI	1863	1871
606	Francis Beresford Wright	1,837	1,911	AI	1870	1871
607	Thomas Wright	1,810	1,877	AI	1853	1871
608	Robert Younge	1,801	1,874	AI	1865	1871
609	Arthur de Zeltner	NaN	NaN	AI	1865	1871

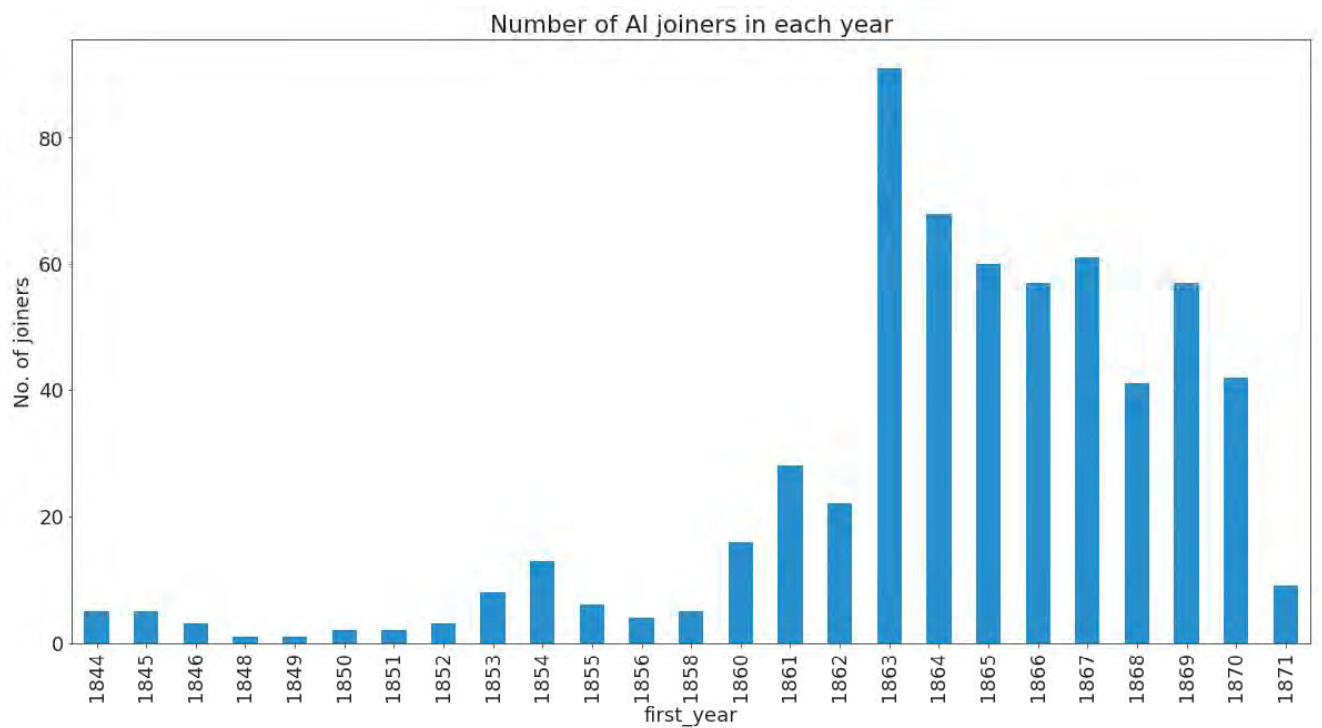
610 rows x 6 columns

```
quakers_ai = pd.read_csv ('vw_4_ceda_membership_quakers_ai2.csv')
quakers_ai
```

	Name	birth_year	death_year	religion_name	ceda_name	person_ce
0	William Spicer Wood	NaN	1902.0	Quaker	AI	
1	Jonathan Hutchinson	1828.0	1913.0	Quaker	AI	
2	Charles Henry Fox	NaN	NaN	Quaker	AI	
3	Robert Nicholas Fowler	1828.0	1891.0	Quaker	AI	
4	Henry Crowley	NaN	1887.0	Quaker	AI	
5	William Bull	1828.0	1902.0	Quaker	AI	
6	Antonio Brady	1811.0	1881.0	Quaker	AI	
7	Edward Backhouse	1808.0	1879.0	Quaker	AI	

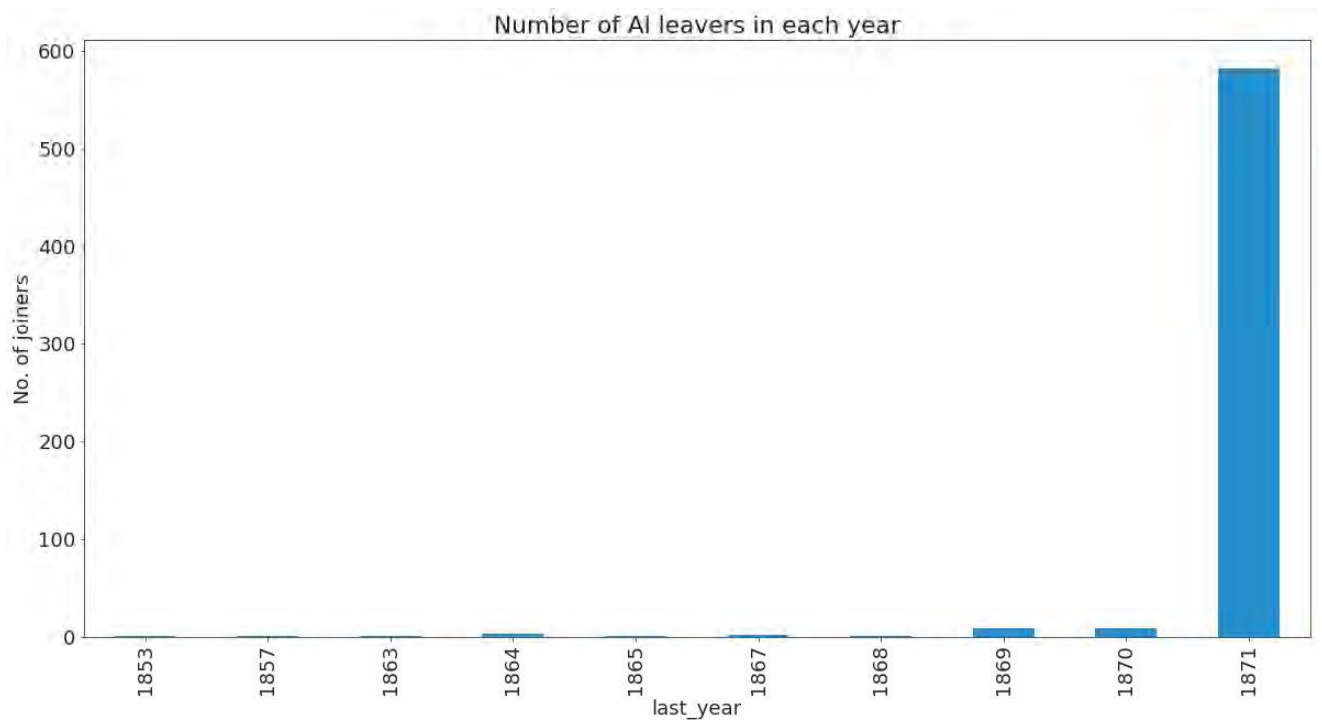
4.23 AI joiners in each year

```
ai.groupby('first_year')['Name'].nunique().plot(kind='bar')
plt.title ("Number of AI joiners in each year")
plt.ylabel ("No. of joiners")
plt.show()
```



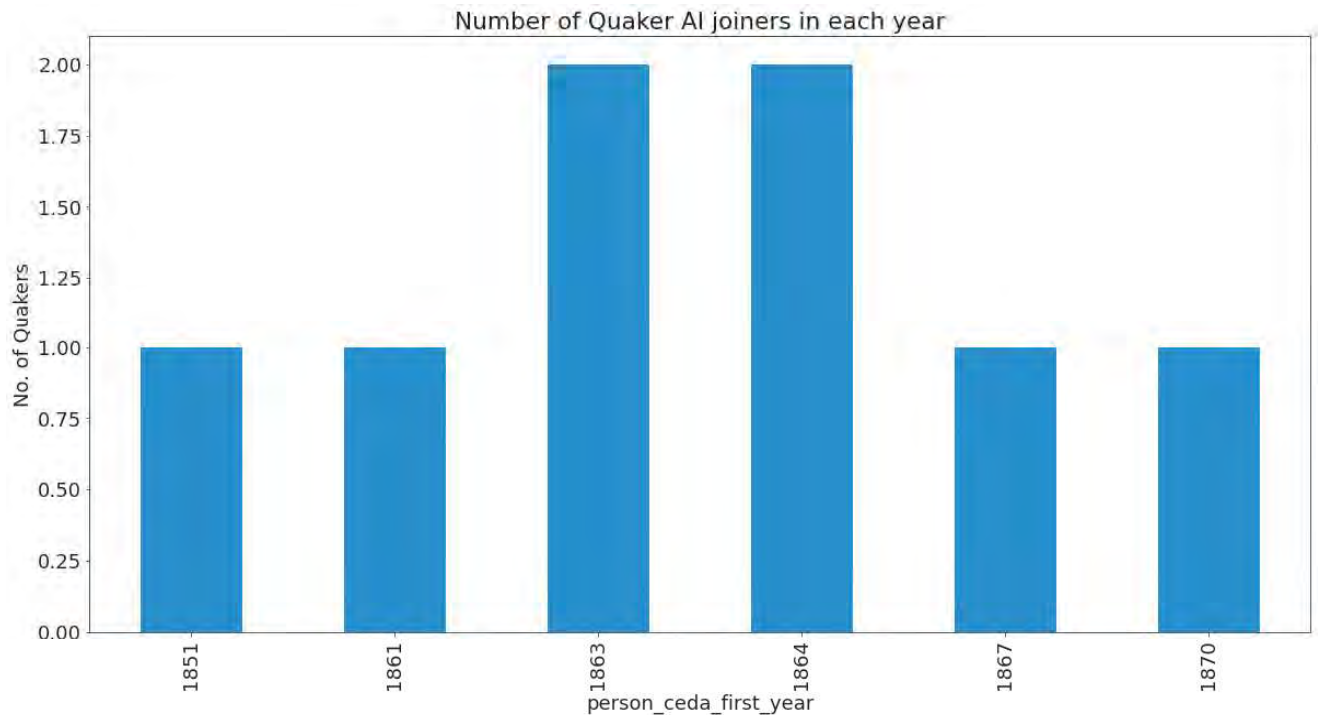
4.24 AI leavers in each year

```
ai.groupby('last_year')['Name'].nunique().plot(kind='bar')
plt.title("Number of AI leavers in each year")
plt.ylabel("No. of joiners")
plt.show()
```



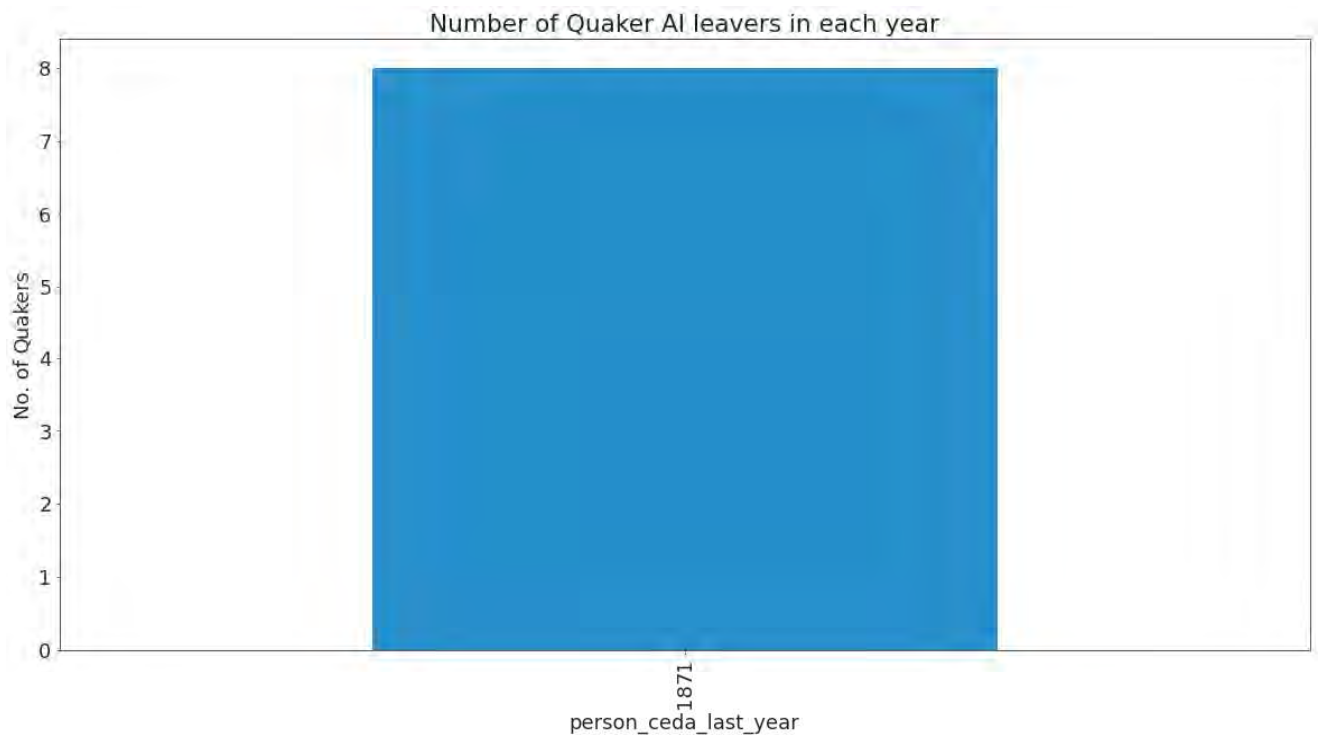
4.25 AI Quaker joiners in each year

```
quakers_ai.groupby('person_ceda_first_year')['Name'].nunique().plot(kind='bar')
plt.title ("Number of Quaker AI joiners in each year")
plt.ylabel ("No. of Quakers")
plt.show()
```



4.26 AI Quaker leavers in each year

```
quakers_ai.groupby('person_ceda_last_year')['Name'].nunique().plot(kind='bar')
plt.title ("Number of Quaker AI leavers in each year")
plt.ylabel ("No. of Quakers")
plt.show()
```

4.27 Duration of AI Quaker memberships



4.28 generate Gexf output file of all CEDA data for Gephi

```
with open('vw_1_person_with_quakers2.csv', 'r') as nodecsv: # Open the Nodes
    nodereader = csv.reader(nodecsv) # Read the csv
    nodes = [n for n in nodereader][1:] # Retrieve the data (using Python list slicing)
                                           # to remove the header row
    node_names = [n[0] for n in nodes] # Get a list of only the node names

with open('vw_hddt_ceda_tuples_attributes2.csv', 'r') as edgescsv: # Open the edges
    edgereader = csv.reader(edgescsv) # Read the csv
    edge_list = list(edgereader) # Convert to list, so can iterate below in

# Create empty arrays to store edge data and edge attribute data
edges = []
edges_attributes = []

# Fill the arrays with data from CSV
for e in edge_list[1:]:
    edges.append(tuple(e[0:2])) # Get the first 2 columns (source, target)
    edges_attributes.append(tuple(e[2:4])) # Get the 3rd and 4th columns

edge_names = [e[0] for e in edges] # Get a list of only the edge names
```

```
print("Nodes length: ", len(node_names))
print("Edges length: ", len(edges))
print("Edges attributes length: ", len(edges_attributes)) # This should be
```

```
Nodes length: 3094
Edges length: 4046
Edges attributes length: 4046
```

```
print("First 5 nodes:", node_names[0:5])
print("First 5 edges:", edges[0:5])
print("First 5 edges attributes:", edges_attributes[0:5])
```

The output will appear below this code cell.

```
First 5 nodes: ['Arthur William A Beckett', 'Andrew Mercer Adam', 'H R Adam']
First 5 edges: [('William Adam', 'ESL'), ('William (1) Adams', 'ESL'), ('Wil
First 5 edges attributes: [('1844', '1844'), ('1844', '1844'), ('1858', '187
```

```
G = nx.Graph()
G.add_nodes_from(node_names)
G.add_edges_from(edges)
print(nx.info(G))
```

```
Name:
Type: Graph
Number of nodes: 3100
Number of edges: 4021
Average degree: 2.5942
```

Nodes

```
birth_year_dict = {}
death_year_dict = {}
religion_id_dict = {}
```

Edges

```
first_year_dict = {}
last_year_dict = {}
```

```

for node in nodes: # Loop through the list, one row at a time

    birth_year_dict [node[0]] = node[1]
    death_year_dict [node[0]] = node[2]
    religion_id_dict[node[0]] = node[3]

```

```

for i, edge in enumerate(edges): # Loop through the list, one row at a time
    first_year_dict [(edge[0], edge[1])] = edges_attributes[i][0]
    last_year_dict [(edge[0], edge[1])] = edges_attributes[i][1]

```

```

# print(religion_id_dict)# list Source, target and first_year (all records).
# print(len(religion_id_dict))# At the end of the file print a count of all
# print (religion_id_dict)

```

```

# Nodes
nx.set_node_attributes(G, birth_year_dict, 'birth_year')
nx.set_node_attributes(G, death_year_dict, 'death_year')
nx.set_node_attributes(G, religion_id_dict, 'religion_id')

# Edges
nx.set_edge_attributes(G, first_year_dict, 'first_year')
nx.set_edge_attributes(G, last_year_dict, 'last_year')

```

```

#for n in G.nodes(): # Loop through every node, in our data "n" will be the
#print(n, G.nodes[n]['birth_year']) # Access every node by its name, and the

```

```

nx.write_gexf(G, 'ceda_all_data_dyn_edges.gexf')

```

P7 Chapter 5a Thomas Hodgkin MD's networks - Part one

file_name: jnb_hddt_laidlaw

5.1 Protecting the Empire's Humanity: Thomas Hodgkin and British Colonial Activism 1830 - 1870 (Zoë Laidlaw 2021)

This 'HDDT - JNB' exercise analyses the persons listed in the index to Protecting the Empire's Humanity (Laidlaw 2021), (PEH), and correspondence received by and correspondence sent to Thomas Hodgkin MD 1799 - 1861, in the indexes to the Wellcome Inst., Hodgkin Family Archives. Selected persons from both archives are then compared with the CEDA database persons. The relationship between the three datasets is then examined to determine the extent to which the networks overlap or fit together; if a Hodgkin political activist network, then emerges, how central is that network to the activism of the wider CEDA network and do insights emerge that might indicate who the key influencers in the network might be?

5.2 GitHub

Make a private GitHub repository for the exercise and clone it to the University of Birmingham secure server space allocated for this project.

[KelvinBeerJones/jnb_hddt_laidlaw](https://github.com/KelvinBeerJones/jnb_hddt_laidlaw) cloned to: [http://localhost:8888/tree/OneDrive-20University of Birmingham/HDDT/jnb_project_containers/jnb_hddt_laidlaw](http://localhost:8888/tree/OneDrive-20University%20of%20Birmingham/HDDT/jnb_project_containers/jnb_hddt_laidlaw)

5.3 Call up the python packages needed to perform the analysis

1. Pandas, numpy and pyplotlib, which we will use to create tables and charts in the Workbook.
2. Plot.rc to specify the dimensions for all imported images (this keeps images to a uniform size and shape).
3. Itemgetter, NetworkX and nbconvert to create a Gexf file for Gephi, which is used to generate visualisation graph files and to enable visual analysis of the social networks to take place.
4. A csv reader to extract the selected sqlite database data from the curated views.

```

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
plt.rc('figure', figsize=(20, 10))
from IPython.display import set_matplotlib_formats
set_matplotlib_formats('png', 'pdf')
from operator import itemgetter
import networkx as nx
from networkx.algorithms import community

#This part of networkx,
# for community detection, needs to be imported separately.
import nbconvert
import csv

# 

```

```

-----
ModuleNotFoundError                                Traceback (most recent call last)
Cell In[1], line 3
      1 import pandas as pd
      2 import numpy as np
---->  3 import matplotlib.pyplot as plt
      4 plt.rc('figure', figsize=(20, 10))
      5 from IPython.display import set_matplotlib_formats

ModuleNotFoundError: No module named 'matplotlib'

```

5.4 Call up the csv files from the SQL db and prepare data for Gephi

```

# produce a 'names' file of nodes and a 'tuples' file of edges_attributes
# to generate the files need to produce GefX files for Gephi.

person_names = pd.read_csv ('vw_hddt_person_table.csv')

person_data_source = pd.read_csv ('vw_hddt_person_with_data_source.csv')

# Use these csv files in the 'with open' statements below
# to generate locations.gexf

names = pd.read_csv ('vw_hddt_person_name.csv')# For nodes csv
tuples = pd.read_csv ('vw_hddt_ceda_tuples.csv')# For edges.csv

with open('vw_hddt_person_name.csv', 'r') as nodecsv:

# Open the Nodes csv file
    nodereader = csv.reader(nodecsv)

# Read the csv
    nodes = [n for n in nodereader][1:]

# Retrieve the data (using Python list comprehension and list slicing
# to remove the header row

    node_names = [n[0] for n in nodes]

# Get a list of only the node names

with open('vw_hddt_ceda_tuples.csv', 'r') as edgecsv:

# Open the file

    edgereader = csv.reader(edgecsv)
# Read the csv

    edge_list = list(edgereader)

# Convert to list, so can iterate below in for loop

# Create empty arrays to store edge data and edge attribute data

edges = []
edges_attributes = []

# Fill the arrays with data from CSV

for e in edge_list[1:]:
    edges.append(tuple(e[0:2]))

# Get the first 2 columns (source, target) and add to array
# not used this time. edges_attributes.append(tuple(e[2:4]))
# Get the 3rd and 4th columns (first_year, last_year) and add to array

edge_names = [e[0] for e in edges] # Get a list of only the edge names

```

5.5 Introduction to the exercise – Part One

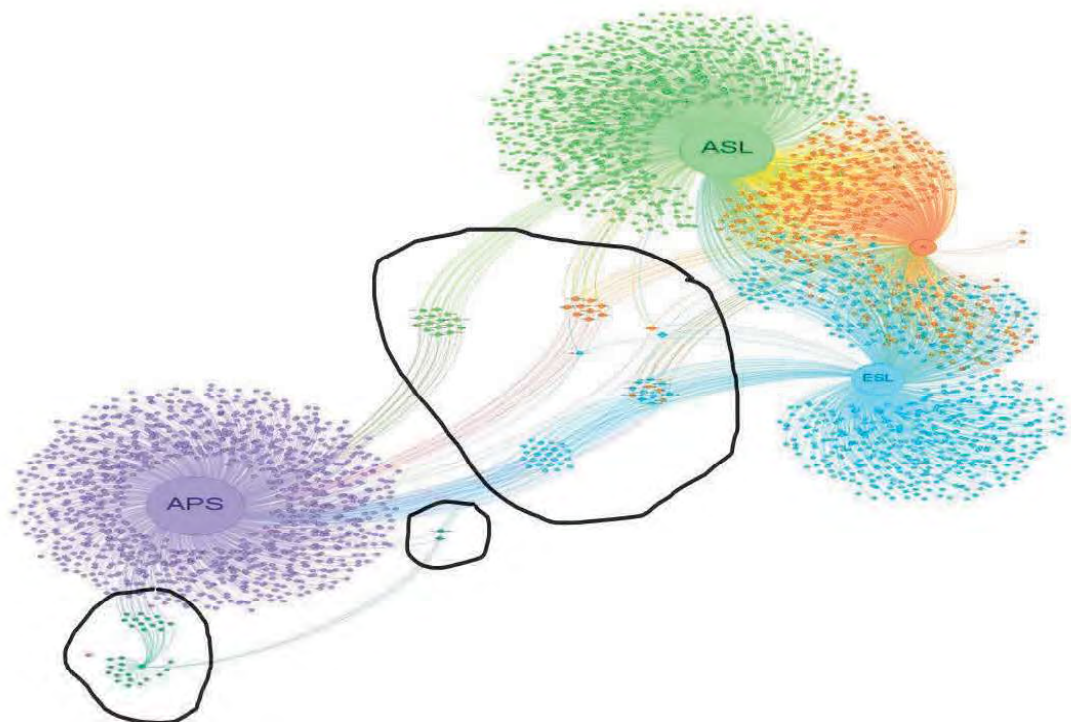
Laidlaw in the introduction to PEH sets out the importance of the social networks of Dr Thomas Hodgkin MD, “The roots of this book lie in the personal correspondence of the Quaker, scientist, and activist, Dr Thomas Hodgkin”, and “The exploration of Britain’s imperial history presented in this book is profoundly shaped by Thomas Hodgkin’s personal papers, today housed in the Wellcome library in London, and what they reveal of his philanthropic, medical, and scientific interests and networks. The volume and the breadth of that archive has allowed me to trace and assess a wide array of influences on “imperial humanitarianism” (2021, 3). Laidlaw also says that, Hodgkin’s personal archive at the Wellcome Institute constitutes The “backbone” of the book and that “Its study reveals 50 years’ worth of unlikely connections, sustained relationships, and closely argued, if sometimes contradictory, cases” (2021, 7).

Chapter 5 of the thesis argues that by only describing them in narratives it is difficult to ‘see’ the people who formed the extensive social networks in which Thomas Hodgkin MD operated, how many there were and how they relate to each other. And so, PEH, while discussing at length the efforts of Hodgkin’s social networks to relieve the plight of aborigines throughout the British Empire, does not make those networks visible or make them available to scrutiny. This Historical Data Analysis (HDA) fully addresses those two needs and presents new insights about Hodgkin’s networks that are difficult to achieve without an HDDT.

5.6 The 3094 members of the CEDA before the exercise

Before the Laidlaw exercise is performed, we can see in the Gephi graph the CEDA network using the Force Atlas algorithm. Far left and at the bottom we can see the QCA and those members of the QCA who led in the formation of the APS (which appears in purple). Two members of the QCA who join the APS also join with non QCA members of the APS in joining the ESL in 1843. These two are William Allen (who would die that year) and Thomas Hodgkin MD. In connecting up the QCA, the APS and the ESL these two stand out from the other members of the QCA, (half of whom go on to join the APS) but participate no further in network building. There is no discernible network centred on Hodgkin alone. Thomas Hodgkin’s personal network does not appear as a group node. The memberships of the ESL, the Anthropological Society of London (ASL) and Anthropological Institute (AI) are frequently shared. The ESL splits into two groups each of roughly equal size, those who are members of only the ESL and those members of the ESL

who also share membership with the AI and ASL. The ASL (formed last in 1863) draws its membership from both the ESL and the AI. It is the ASL that will provide the bulk of the first memberships of the RAI. We can see 'ringed' the small groups of key influencers who network between the CEDA groups.



Society	abv.	Dates	Colour
Quaker Committee on the Aborigines*	QCA	1832/37 - 1846	Dark green
Aborigines Protection Society	APS	1837 - 1919	Purple
Ethnological Society of London	ESL	1843 - 1871	Blue
Anthropological Institute	AI	1843 - 1871	Orange
Anthropological Society of London	ASL	1863 - 1871	Green

5.7 Mods to db to facilitate this exercise - ZOE and WEL

(1) Persons in the index to PEH, and the members of the CEDA

All the person names were extracted from the PEH index (Laidlaw 2021, 359). Of the 290 person names in the index 108 were found to be already present in the HTTD CEDA dataset. The remaining 182 do not appear in the HDDT CEDA dataset, some of those indexed in PEH will not be members of Hodgkin’s support network, but rather persons that Laidlaw has referenced in PEH for other reasons (for example King William IV, Queen Victoria, and many colonial officers of the Crown).

all persons PEH	PEH persons CEDA	PEH persons not CEDA
290	108	182

Note: At least 15 persons amongst the 182 non-CEDA persons appearing in the index to PEH could be family members of persons already captured in the HDDT - from the awareness of the author of this thesis. Nonetheless the intention here is to discover how the 108 who are recorded in the HDDT database relate both to each other and the wider group of 3000 already in the HDDT, and to disregard the 182 who are not.

(2) Persons in the indexes to WEL, and the members of the CEDA

Because PEH relies heavily on the Hodgkin Family Archive at the Wellcome Institute an analysis of the two indexes to that collection was performed to extract from the index of letters sent and the index of letters received by Thomas Hodgkin MD those persons who appeared in both indexes indicating these persons might have been members of Hodgkin’s network.

In the Wellcome Hodgkin Family archive data indexes there are 107 person Hodgkin writes to and who also write to him, and of these 46 appear in the HDDT database and 61 do not. As in the PEH index exercise above we can reasonably assume both that some of these persons not already in the HDDT may be related to members of persons appearing in the HDDT database; and also, that many will not, for example we can expect to find many

medical related correspondences in an archive collected by a medical history archive, and these are outside the scope of this thesis (which analyses political activism).

all persons WEL	WEL persons CEDA	WEL persons not CEDA
107	46	61

(3) Summary of db mods

Using the platform DBeaver (<https://dbeaver.io/>) the HDDT CEDA database was modified to accommodate the exercise:

1. We added to person_data_source table two new temporary data sources – ZOE and WEL
2. We added to ceda table two new CEDA groups – ZOE and WEL
3. Person_table. We uploaded 182 new records from a csv file of Laidlaw references, and allocated them to data_source = ZOE.
4. Person_table. We uploaded 61 new records from a csv file of Laidlaw references, and allocated them to data_source = WEL.
5. We updated m2m_person_ceda table to allocate 108 persons to a new CEDA group = ZOE
6. We updated m2m_person_ceda table to allocate 46 persons to a new CEDA group = WEL

Once the data from both ZOE and WEL had been added to the HDDT database the network analysis could be performed and a report in the form of a Jupyter Notebook made.

5.8 Assess persons in the index to PEH who are members of the CEDA.

The Index to PEH lists 108 persons who already present in the HDDT CEDA db. They were Laidlaw allocated to a 'dummy' CEDA group (table CEDA, Target = 'ZOE'). This enables the visualisation of Laidlaw's Hodgkin network and shows its relationships to the original HDDT CEDA groups.

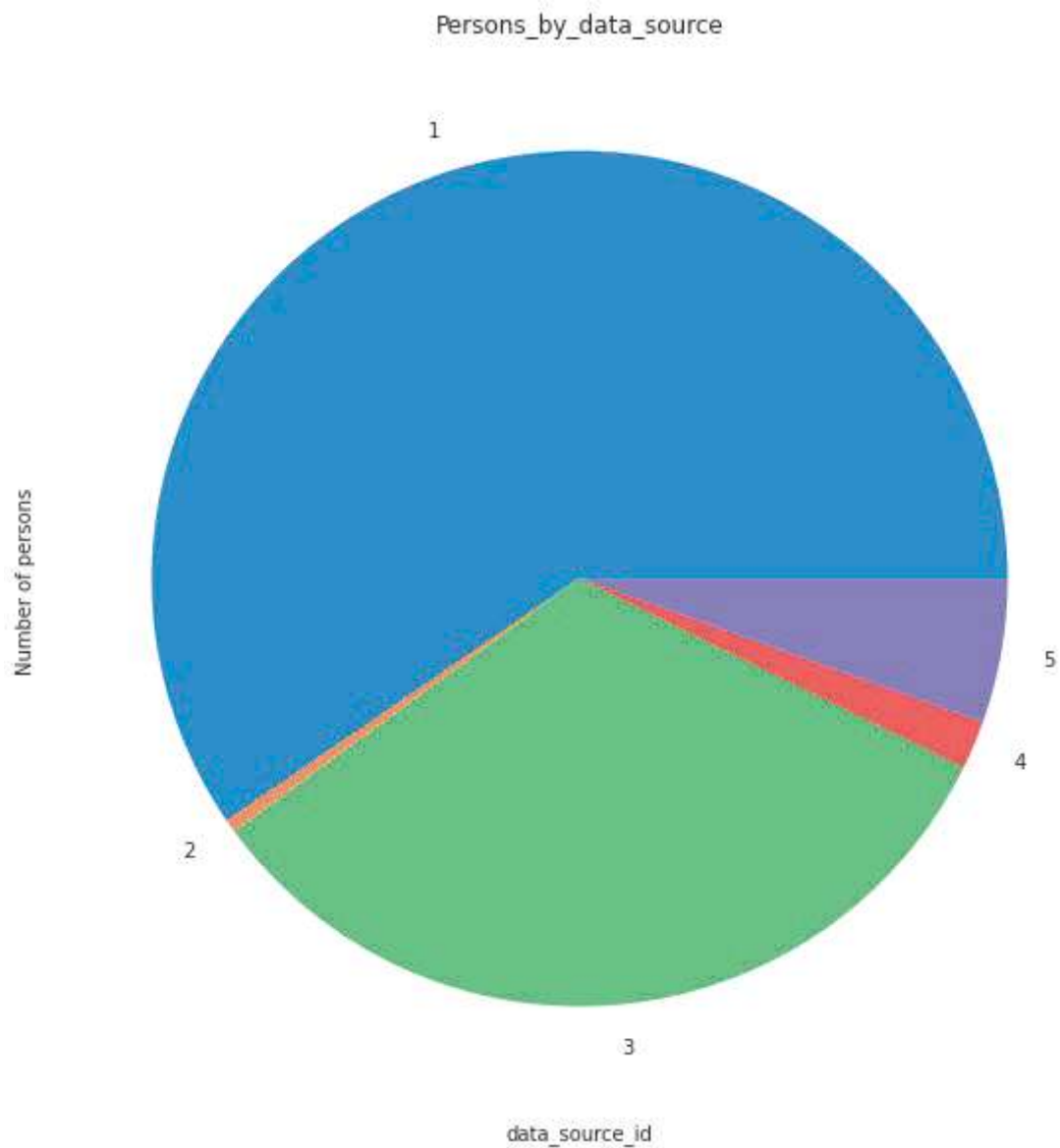
The indexes to WEL list 46 persons who are already present in the HDDT CEDA db. They were allocated to a 'dummy' CEDA group (table CEDA, Target = 'WEL'). This enables the visualisation of The Wellcome Inst., Hodgkin Family Archive Thomas Hodgkin MD network and shows its relationships to the original HDDT CEDA groups.

The 3337 persons in the CEDA db., make up 3892 memberships of the original CEDA (QCA, APS, ESL, AI, and the APS) The above 'new' CEDA were set up in the SQL db., to facilitate this exercise.

Original CEDA	ZOE	WEL	Total new
3892	108	46	4046

5.9 CEDA members compared to non-CEDA others in PEH index

```
person_names.groupby('data_source_id')['Name'].nunique().plot(kind='pie')
plt.title ("Persons_by_data_source")
plt.xlabel ("data_source_id")
plt.ylabel ("Number of persons")
plt.show()
```



data_source_id	Source
1	RAI
2	QCA
3	APS
4	WEL
5	ZOE

The persons from PEH Index and the Welcome Inst., indexes who are not members of a CEDA appear here. They are disregarded in this exercise because they do not indicate the presence of a social network. They are shown here only for completeness.

Add 182 'ZOE' and 61 'WEL' Non CEDA persons to the SQL db., 'temporarily' solely to visualise the extent of persons not included in this exercise.

5.10 Data verification

In each of Code cells 4 - 11 We call up the tables we have obtained from the db.to view the data. We can confirm (e.g., Code cell 4) that we have selected the correct table or view from the db., and we check that the first 5 and last 5 records have been rendered correctly. We can confirm (e.g. Code cell 5) that the respective table dataframe is formatted as expected. We also check that the number of records equals the number of rows on the data source csv.

Code cell 4

```
person_names
```

	Name	title	gender_id	birth_year	death_year	data_source_id
0	Arthur William A Beckett	NaN	1.0	1844.0	1909.0	1
1	Andrew Mercer Adam	NaN	1.0	NaN	NaN	1
2	H R Adam	NaN	1.0	NaN	NaN	1
3	William Adam	NaN	1.0	NaN	NaN	1
4	Henry John Adams	NaN	1.0	NaN	NaN	1
...
3332	James Wetherall	NaN	NaN	NaN	NaN	5
3333	William Wilberforce	NaN	NaN	NaN	NaN	5
3334	King William IV	NaN	NaN	NaN	NaN	5
3335	x Wiremu Kingi Te rangitake	NaN	NaN	NaN	NaN	5
3336	Johann Wohlers	NaN	NaN	NaN	NaN	5

3337 rows x 6 columns

Code cell 5

```
person_names.info()
```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3337 entries, 0 to 3336
Data columns (total 6 columns):
#   Column                Non-Null Count  Dtype  
---  -
0   Name                   3337 non-null   object  
1   title                  776 non-null    object  
2   gender_id              1988 non-null   float64 
3   birth_year             1003 non-null   float64 
4   death_year             1069 non-null   float64 
5   data_source_id         3337 non-null   int64   
dtypes: float64(3), int64(1), object(2)
memory usage: 156.5+ KB

```

Code cell 6

```
person_data_source
```

	Name	data_source
0	Arthur William A Beckett	RAI
1	Andrew Mercer Adam	RAI
2	H R Adam	RAI
3	William Adam	RAI
4	Henry John Adams	RAI
...
3332	James Wetherall	ZOE
3333	William Wilberforce	ZOE
3334	King William IV	ZOE
3335	x Wiremu Kingi Te rangitake	ZOE
3336	Johann Wohlers	ZOE

3337 rows x 2 columns

Code cell 7

```
person_data_source.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3337 entries, 0 to 3336
Data columns (total 2 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Name             3337 non-null   object
1   data_source      3337 non-null   object
dtypes: object(2)
memory usage: 52.3+ KB
```

Code cell 8

```
names
```

	Name
0	Arthur William A Beckett
1	Andrew Mercer Adam
2	H R Adam
3	William Adam
4	Henry John Adams
...	...
3332	James Wetherall
3333	William Wilberforce
3334	King William IV
3335	x Wiremu Kingi Te rangitake
3336	Johann Wohlers

3337 rows x 1 columns

Code cell 9

```
names.info()
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3337 entries, 0 to 3336
Data columns (total 1 columns):
#   Column  Non-Null Count  Dtype
---  -
0    Name    3337 non-null      object
dtypes: object(1)
memory usage: 26.2+ KB
```

Code cell 10

```
tuples
```

	Source	Target
0	William Adam	ESL
1	William (1) Adams	ESL
2	William (2) Adams	ESL
3	Louis Agassiz	ESL
4	Augustine Aglio	ESL
...
4041	Frederick Cooper	WEL
4042	Henry Christy	WEL
4043	James (1) Backhouse	WEL
4044	William (Capt.) Allen	WEL
4045	William (1) Adams	WEL

4046 rows x 2 columns

Code cell 11

```
tuples.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4046 entries, 0 to 4045
Data columns (total 2 columns):
#   Column   Non-Null Count  Dtype
---  -
0    Source   4046 non-null   object
1    Target   4046 non-null   object
dtypes: object(2)
memory usage: 63.3+ KB
```

5.11 Generate gexf file for Gephi visualisation

Code cell 12 - 13 Check that Gephi 'Nodes' and 'Edges' files agree with 'names' and 'tuples' files

Code cell 12

```
print("Nodes length: ", len(node_names))
print("Edges length: ", len(edges))
```

```
Nodes length: 3337
Edges length: 4046
```

Code cell 13

```
print("First 5 nodes:", node_names[0:5])
print("First 5 edges:", edges[0:5])
```

```
First 5 nodes: ['Arthur William A Beckett', 'Andrew Mercer Adam', 'H R Adam', 'H R Adam', 'H R Adam']
First 5 edges: [('William Adam', 'ESL'), ('William (1) Adams', 'ESL'), ('William (1) Adams', 'ESL'), ('William (1) Adams', 'ESL'), ('William (1) Adams', 'ESL')]
```

Code cell 14 Execute NetworkX function

```
# We use NetworkX to build the graph data into a table
```

```
G = nx.Graph()  
G.add_nodes_from(node_names)  
G.add_edges_from(edges)  
print(nx.info(G))
```

```
Name:  
Type: Graph  
Number of nodes: 3344  
Number of edges: 4046  
Average degree: 2.4199
```

The number of nodes here is 3337 persons plus 7 societies = 3344 (the additional 7 nodes are the CEDA names (CQA, APS, ESL, AI, ASL, ZOE and WEL). Gephi will produce a 'bigraph' of the selected data and a bigraph is a graph where relationships are between individual persons (one node) and membership organisations, many nodes).

Code cell 15 Write the gexf file

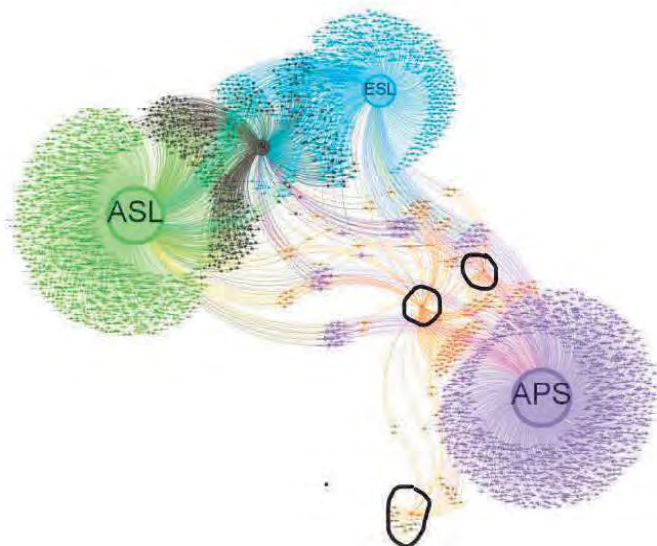
```
# Finally we can write a gexf file which will be placed in the root directory  
# We can then open the file in Gephi and visualize the network.  
  
nx.write_gexf(G, 'ceda_laidlaw.gexf')
```

5.12 Visual analysis of the exercise

We then open the gexf file in Gephi and generate a graph file using the Force Atlas algorithm, the 'network diameter', 'modularity' routines and the 'appearance' routines to produce a suitable graph for analysis. (Running these routines in Gephi allows graph display and placement, colour and size of nodes, modularity to identify clearly the large society groups and betweenness centrality to reveal individuals and the smaller groups who play important roles linking groups together. It should be noted that the visualisation and all of its topography are the result of running the above algorithms, it is not manually arranged on the page! Finally we save the graph in Gephi as a 'project file' to the JNB container for this exercise, and we also produce PNG files of the graph and selected areas. Finally we can import the png images of the Gephi graph file to this JNB to illustrate the network analysis performed in the Gephi platform.

5.13 The CEDA social network including ZOE and WEL

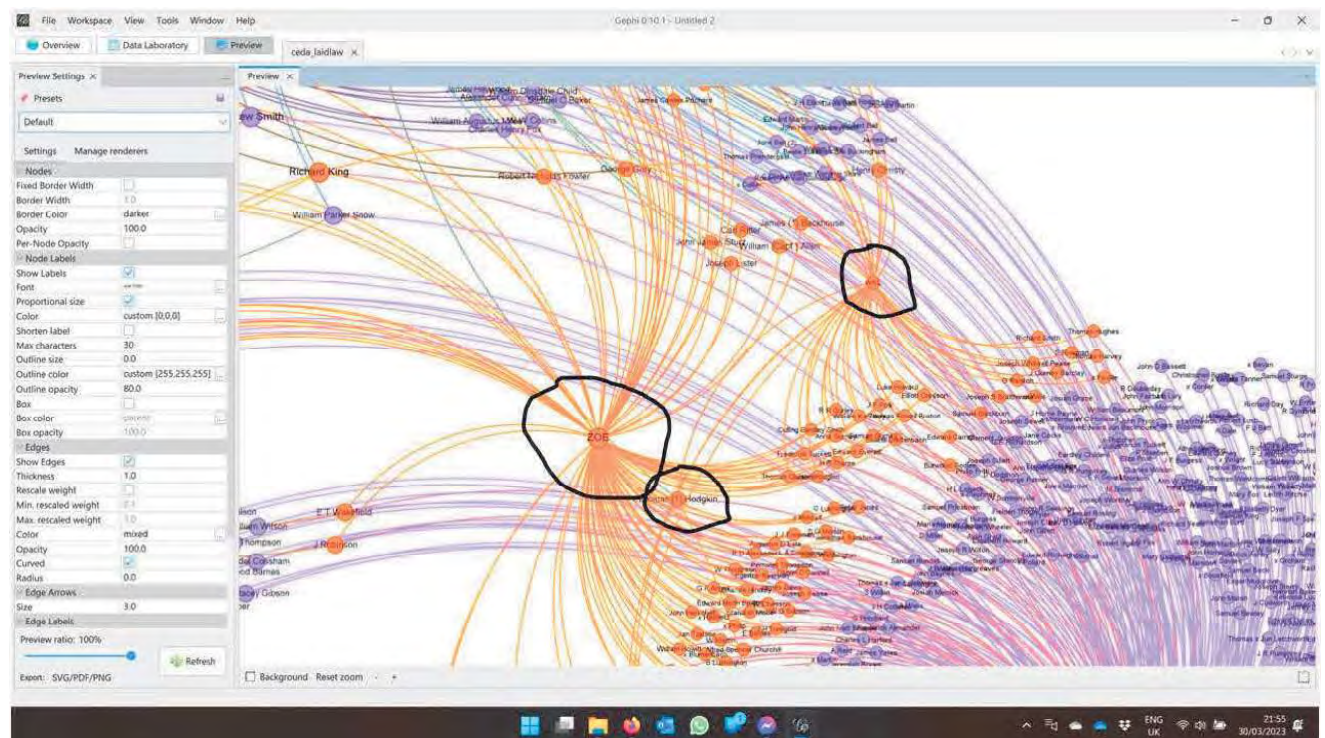
We have created two 'dummy' groups (ZOE and WEL) to show the persons who are members of the original CEDA and who appear in the index of PEH as Zoe, and those from the Wellcome Inst., as WEL. They appear in this graph in orange. Also in orange is the QCA group. It is helpful to show QCA alongside the two dummy groups because Thomas Hodgkin MD begins his political work in the QCA (as discussed in PEH), and all three groups are arguably Thomas Hodgkin MD groups. ZOE is centre, WEL to the right and QCA to the left. We can see that the dummy groups are centred in the Force Atlas graph and that they connect up all of the main CEDA groups. This is a visible confirmation that the Thomas Hodgkin MD network referenced in PEH is well placed and well connected. The presence of the dummy groups brings Thomas Hodgkin MD to the very centre of the graph, this is impressive given that the entire population is 3337 persons. But it is important to note that Hodgkin does not 'sit' within the community referenced by Laidlaw or that suggested by the WEL analysis, he sits to one side because as much as he is attracted to Laidlaw's and the WEL groups he is 'pulled' away from them because of his connections to the QCA and ESL.



Society	abv.	Dates	Colour
Quaker Committee on the Aborigines, Protecting the Empire's Humanity and Welcome Inst.,	QCA, PEH, WEL	1832/37 - 1846	Orange
Aborigines Protection Society	APS	1837 - 1919	Purple
Ethnological Society of London	ESL	1843 - 1871	Blue
Anthropological Society of London	ASL	1863 - 1871	Green
Anthropological Institute	AI	1843 - 1871	grey

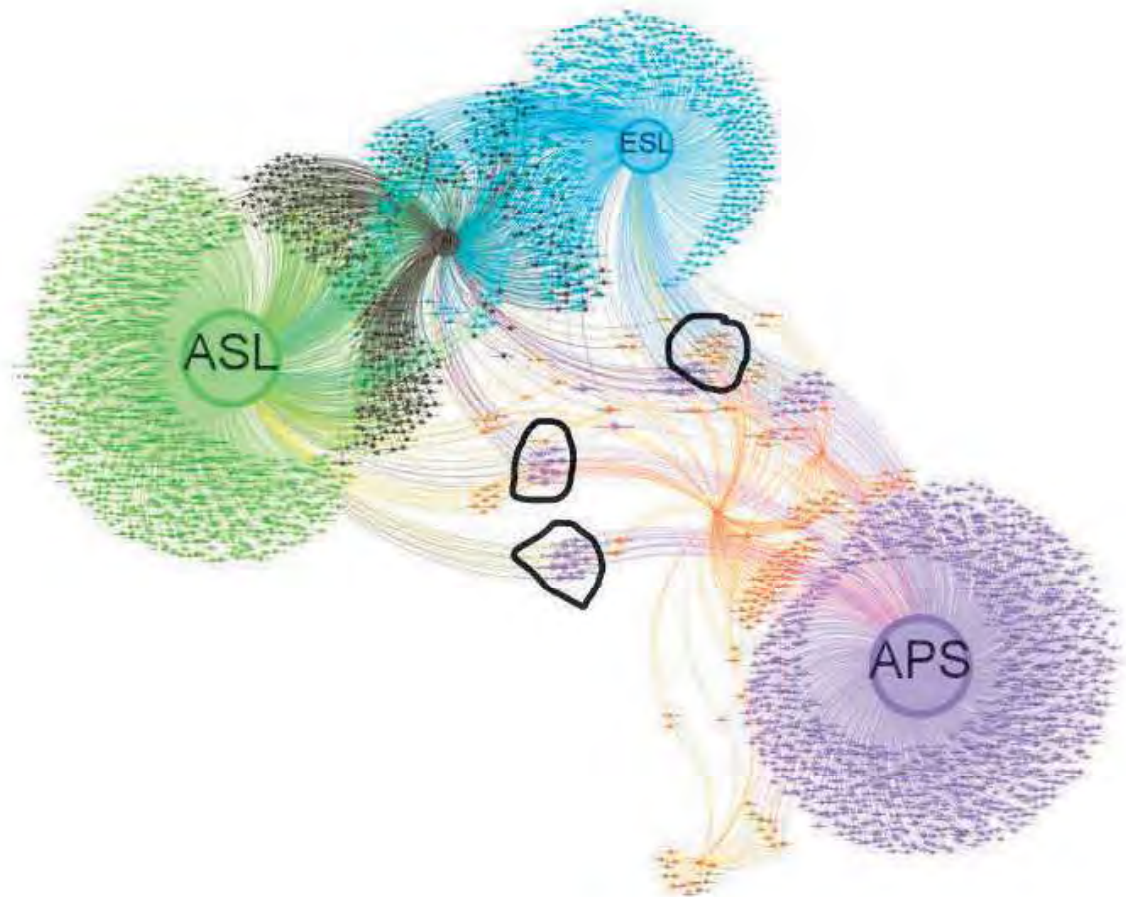
Note: Gephi allows detailed examination of this network

5.14 Zooming in to show the network in detail



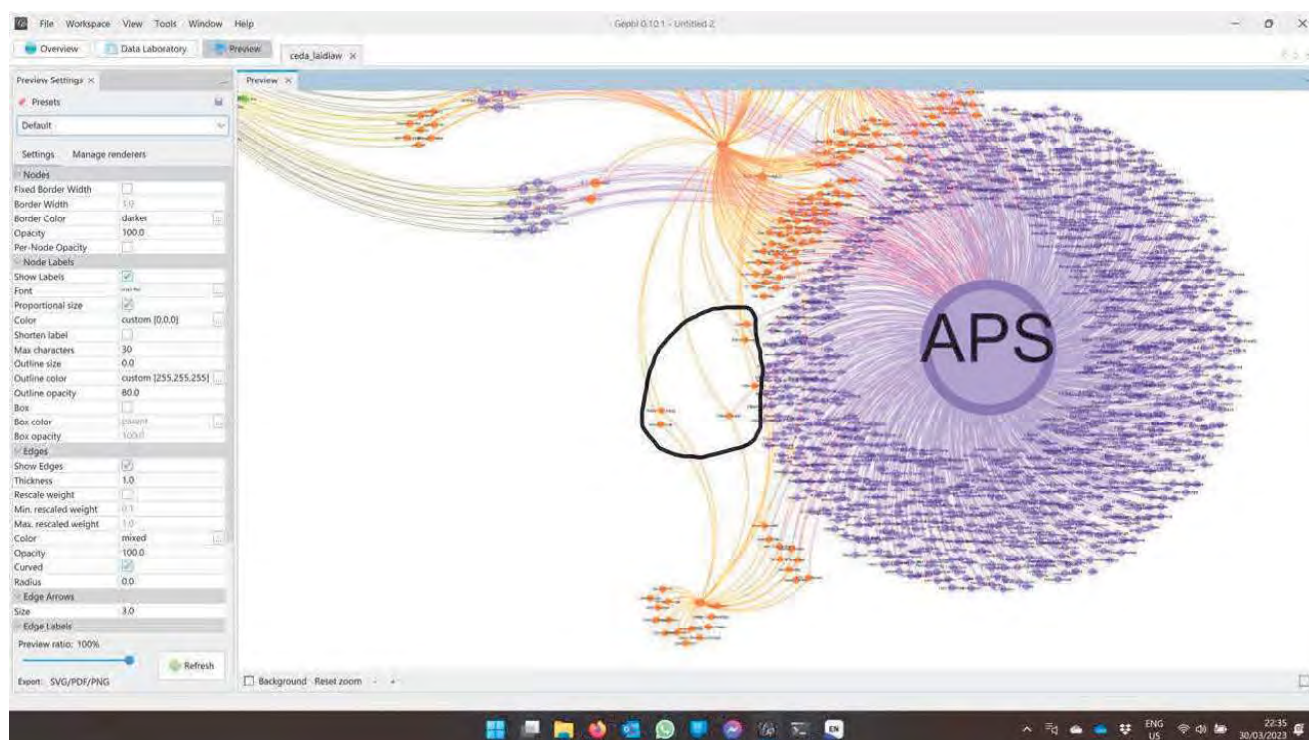
5.15 Other groupings emerge

Other smaller groups that liaise between Hodgkin and the CEDA also become visible in the graph, and these are worthy of further study.



5.16 Quaker roles emerge in detail

PEH also references 6 (of the 15) Quakers who were members of the QCA - Josiah and William Forster, Robert Howard, Peter Bedford, Joseph Sturge and Robert Alsop (Jun) and they can be seen here networking between the QCA, APS, ZOE and Thomas Hodgkin MD.



5.17 Conclusions

This exercise has shown that although Thomas Hodgkin's networks appear only in the index in PEH and in the Wellcome Inst., Hodgkin Archive, in neither of these sources can political activist relationships be easily deduced but, by using data analysis and data visualisation technology the networks argued in PEH and hidden in WEL can be shown and analysed very clearly.

The HDDT CEDA db., comprises the 3000 memberships of CEDA societies concerned firstly with the plight of aborigines, but then quickly afterwards with institution building in the science of ethnology and anthropology and within which Thomas Hodgkin was a key influencer. Placing PEH's Hodgkin networks alongside the HDDT networks, a much fuller view of community's political activism is obtained, where the centrality of Hodgkin's networks to the work of all of the CEDA networks is evident. Laidlaw in PEH has in referencing over 100 key networkers, identified those who have close working connections with Hodgkin, and these networks are central to the workings of the greater networks of the CEDA. Other smaller embedded networking groups that liaise between Hodgkin and the CEDA have also been revealed, and these are worthy of further detailed study.

Although PEH offers a literary description of the networks of Hodgkin, Hodgkin himself does not get subsumed by the Gephi graph algorithms into ZOE or WEL, instead he stands aside. This is because, as we can see, his relationships with Quakers and the APS are strong enough to resist the attraction of his relationships with ZOE or WEL networks. This finding is important because it gives good reason to examine Hodgkin's relationships

further with both QCA and the APS (and the ESL). Six Quakers who are key networkers between Hodgkin's concern for the plight of aborigines and his institution building in the science of anthropology also can be identified and their presence will help to shape the next exercise. The exercise confirms and supports the centrality and importance of Hodgkin's networks as set out in the index to PEH and invites further scrutiny of those networks because key individuals and small clusters of persons emerge from the visual analysis.

5.18 Modifications to the database for Part Two

The 'dummy' groups ZOE and WEL, after being scrutinised using a Gephi graph file, add richness and insight into the workings of the CEDA networks and the role of Thomas Hodgkin within those networks. The role of Quakers within these networks is also indicated. Because the new data in ZOE and WEL does not include visual anomalies or reveal a large number of outlying persons (who would add no value to a study of person-to-person networks) we can be confident in now modifying the CEDA database permanently by merging them both into a new CEDA group called HOD. This part of the exercise will be performed in Part Two.

5.19 Github upload

We can now update GitHub to pass the exercise and all its resources to the project repo (see step 1) . This enables the entire exercise to be both scrutinised by others and replicated elsewhere. (When the exercise is completed and audited, and all copyright issue resolved the repo can be made public).

End of laidlaw (part one)

P7 Chapter 5b Thomas Hodgkin's MD networks - Part two

File name: jnb_hddt_laidlaw2

Protecting the Empire's Humanity (PEH): Thomas Hodgkin and British Colonial Activism 1830 - 1870 (Zoë Laidlaw 2021)

5.20 Preparation

In Part Two of the exercise we modify the HDDT database to accept the data extracted from PEH and WEL, labelling it as a new CEDA called 'HOD' to show Thomas Hodgkin MD's personal network. (note: This is his political network that Laidlaw observed in archival research as primarily supporting Hodgkin's work to relieve the plight of Aborigines. It does not include his 'medical' or 'scientific' networks.)

5.21 GitHub

Make a private GitHub repository for the exercise and clone it to the University of Birmingham secure server space allocated for this project.

 [KelvinBeerJones/jnb_laidlaw2](https://github.com/KelvinBeerJones/jnb_laidlaw2) cloned to this container

5.22 Call up the Python packages needed to perform the analysis

1. Pandas, numpy and pyplotlib, - used to create tables and charts in the Workbook.
2. Plot.rc - to specify the dimensions for all imported images (this keeps images to a uniform size and shape).
3. Itemgetter, NetworkX and nbconvert - to create a Gexf file for Gephi, which is used to generate visualisations and to perform visual analysis of the social networks.
4. csv reader - to extract the selected sqlite database data from the selected db., views.

```
# First we call up the python packages we need to perform the analysis:
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
plt.rc('figure', figsize=(20, 10))
from IPython.display import set_matplotlib_formats
set_matplotlib_formats('png', 'pdf')
from operator import itemgetter
import networkx as nx
from networkx.algorithms import community

#This part of networkx,
# for community detection, needs to be imported separately.
import nbconvert
import csv

# 
```

```
-----
ModuleNotFoundError                                Traceback (most recent call last)
Cell In[1], line 5
      3 import pandas as pd
      4 import numpy as np
----> 5 import matplotlib.pyplot as plt
      6 plt.rc('figure', figsize=(20, 10))
      7 from IPython.display import set_matplotlib_formats

ModuleNotFoundError: No module named 'matplotlib'
```

5.23 call up the csv files and prepare data for Gephi visualisation

```

person_names = pd.read_csv ('vw_hddt_person_table2.csv')
hod = pd.read_csv ('laidlaw_hod.csv')

names2 = pd.read_csv ('vw_hddt_person_name2.csv')# For nodes csv
tuples2 = pd.read_csv ('vw_hddt_ceda_tuples2.csv')# For edges.csv

with open('vw_hddt_person_name2.csv', 'r') as nodecsv:

# Open the Nodes csv file
    nodereader = csv.reader(nodecsv)

# Read the csv
    nodes = [n for n in nodereader][1:]

# Retrieve the data (using Python list comprehension and list slicing
# to remove the header row

    node_names = [n[0] for n in nodes]

# Get a list of only the node names

with open('vw_hddt_ceda_tuples2.csv', 'r') as edgecsv:

# Open the file

    edgereader = csv.reader(edgecsv)
# Read the csv

    edge_list = list(edgereader)

# Convert to list, so can iterate below in for loop

# Create empty arrays to store edge data and edge attribute data

edges = []
edges_attributes = []

# Fill the arrays with data from CSV

for e in edge_list[1:]:
    edges.append(tuple(e[0:2]))

# Get the first 2 columns (source, target) and add to array
# not used this time. edges_attributes.append(tuple(e[2:4]))
# Get the 3rd and 4th columns (first_year, last_year) and add to array

edge_names = [e[0] for e in edges] # Get a list of only the edge names

```

5.24 Introduction to the exercise - Part Two

As a result of the exercise performed in Laidlaw (Part One) we accept that a new CEDA can be created called HOD to represent the personal political network of Thomas Hodgkin MD as extracted from the Index to PEH and the indexes to the Wellcome Inst., Hodgkin

Family Archive. All of the members of this new CEDA are also already recorded in the HTTD database as members of at least one of the original CEDA. In this part of the exercise we amend the HTTD database to include the CEDA group HOD.

Laidlaw's research into the Wellcome Inst., Hodgkin Collection reveals an important and relevant social network amongst the HDDT CEDA persons – based on Laidlaw's close study of Thomas Hodgkin's personal correspondence (where the receiver of correspondence also writes to him). We select some of Laidlaw's person references because they already appear in the HDDT CEDA database, and they complement the data initially collected for this thesis because they provide new information about existing HDDT persons and networking.

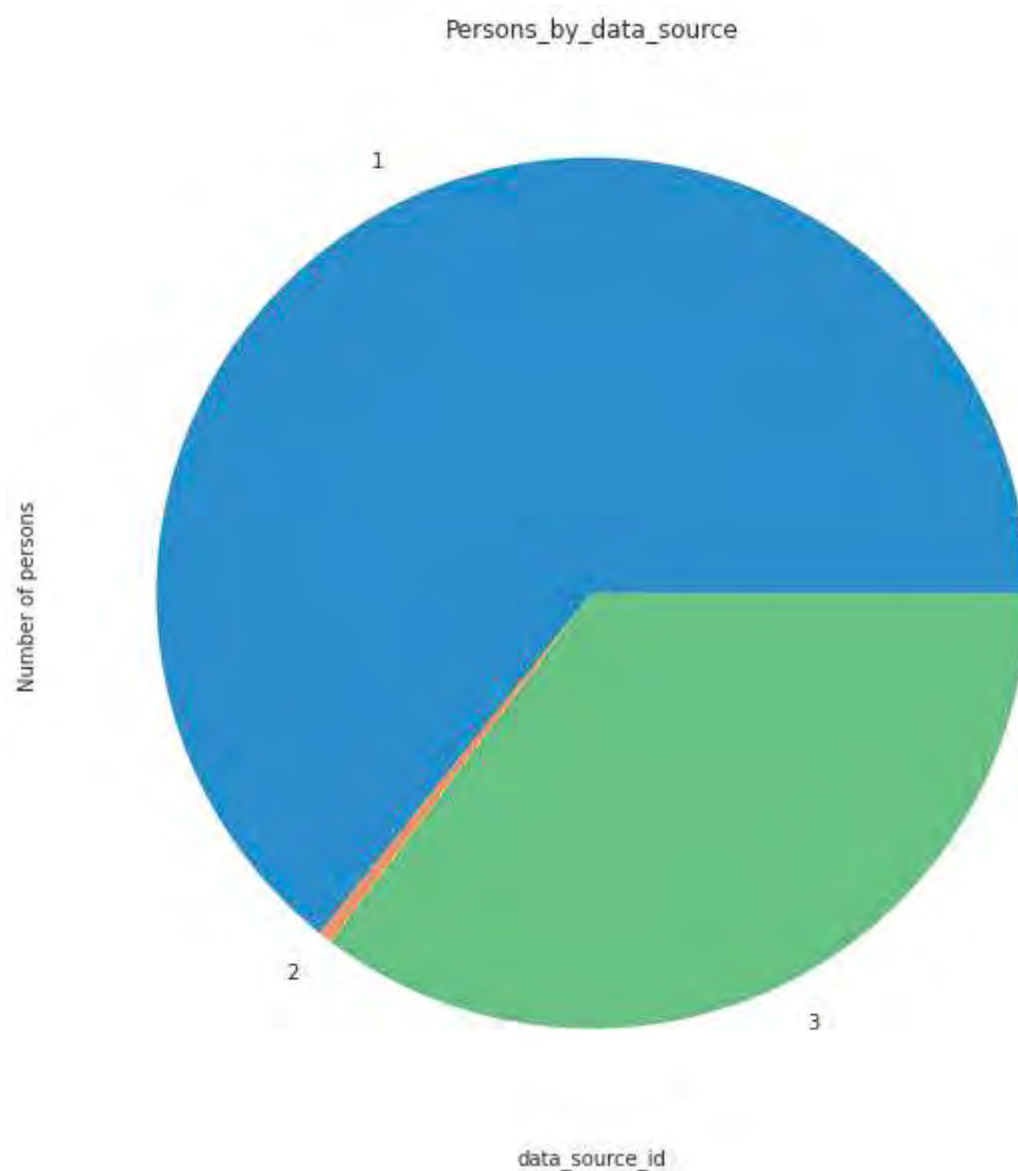
This necessitates amendments to the SQLite database and the production of another (this) Jupyter Notebook made to examine and verify the deepened HDDT CEDA networking in the HDDT as a whole after the Laidlaw HOD modifications had been made.

5.25 Data verification

The members of the Centres for the Emergence of the Discipline of Anthropology in Britain (CEDA) after accepting the new Laidlaw Exercise (Part One) data assigning persons to a new CEDA called HOD. We show that the data sources 'WEL' and 'ZOE' used in Laidlaw Exercise (part One) have been removed from the database. (They were temporarily placed in the HDDT solely as 'dummies' to show the number of persons referenced in PEH and WEL with two way correspondence, but disregarded by the Laidlaw (Part One) exercise because these person are not present in the original CEDA).

5.26 Person table after modification

```
person_names.groupby('data_source_id')['Name'].nunique().plot(kind='pie')
plt.title ("Persons_by_data_source")
plt.xlabel ("data_source_id")
plt.ylabel ("Number of persons")
plt.show()
```



data_source_id	Source
1	RAI
2	QCA
3	APS

Code cell 22 - Person table data

```
person_names
```

	Name	title	gender_id	birth_year	death_year	data_source_id	
0	Arthur William A Beckett	NaN	1.0	1844.0	1909.0	1	S , S. Ja
1	Andrew Mercer Adam	NaN	1.0	NaN	NaN	1	Linc
2	H R Adam	NaN	1.0	NaN	NaN	1	Ca
3	William Adam	NaN	1.0	NaN	NaN	1	
4	Henry John Adams	NaN	1.0	NaN	NaN	1	14 Sc
...	
3089	x Wright	Rev Dr	NaN	NaN	NaN	3	
3090	W Wrigley	NaN	NaN	NaN	NaN	3	
3091	James Yates	Rev	NaN	NaN	NaN	3	
3092	John Young	NaN	NaN	NaN	NaN	3	
3093	Thomas Zachary	NaN	NaN	NaN	NaN	3	

3094 rows x 7 columns



5.27 Persons who are members of the new HOD CEDA

hod

	person_id	Name	birth_year	death_year	Target
0	3386	John Washington	NaN	NaN	HOD
1	3371	Jan Tzatzoe	NaN	NaN	HOD
2	3366	J H Tredgold	NaN	NaN	HOD
3	3359	H B Thorpe	NaN	NaN	HOD
4	3357	Perronet Thompson	NaN	NaN	HOD
...
149	465	Frederick Cooper	NaN	NaN	HOD
150	403	Henry Christy	1,810	1,865	HOD
151	81	James (1) Backhouse	1,794	1,869	HOD
152	28	William (Capt.) Allen	NaN	NaN	HOD
153	7	William (1) Adams	NaN	NaN	HOD

154 rows x 5 columns

5.28 Generate gexf file for Gephi visualisation

Code cell 24 - Check volume of data for Gephi visualisation graph

```
print("Nodes length: ", len(node_names))
print("Edges length: ", len(edges))

# not used this time.
print("Edges attributes length: ", len(edges_attributes))

# This should be the same length as edges
```

```
Nodes length: 3094
Edges length: 4046
Edges attributes length: 0
```

Code cell 25 - Check the data quality for Gephi visualisation graph

```
# First check that the data is correctly formatted

print("First 5 nodes:", node_names[0:5])
print("First 5 edges:", edges[0:5])
# not used this time. print("First 5 edges attributes:", edges_attributes[0

# The output will appear below this code cell.
```

```
First 5 nodes: ['Arthur William A Beckett', 'Andrew Mercer Adam', 'H R Adam',
First 5 edges: [('William Adam', 'ESL'), ('William (1) Adams', 'ESL'), ('Wil
```

Code cell 26 - NetworkX function

```
# We use NetworkX to build the graph data into a table

G = nx.Graph()
G.add_nodes_from(node_names)
G.add_edges_from(edges)
print(nx.info(G))
```

```
Name:
Type: Graph
Number of nodes: 3100
Number of edges: 4021
Average degree: 2.5942
```

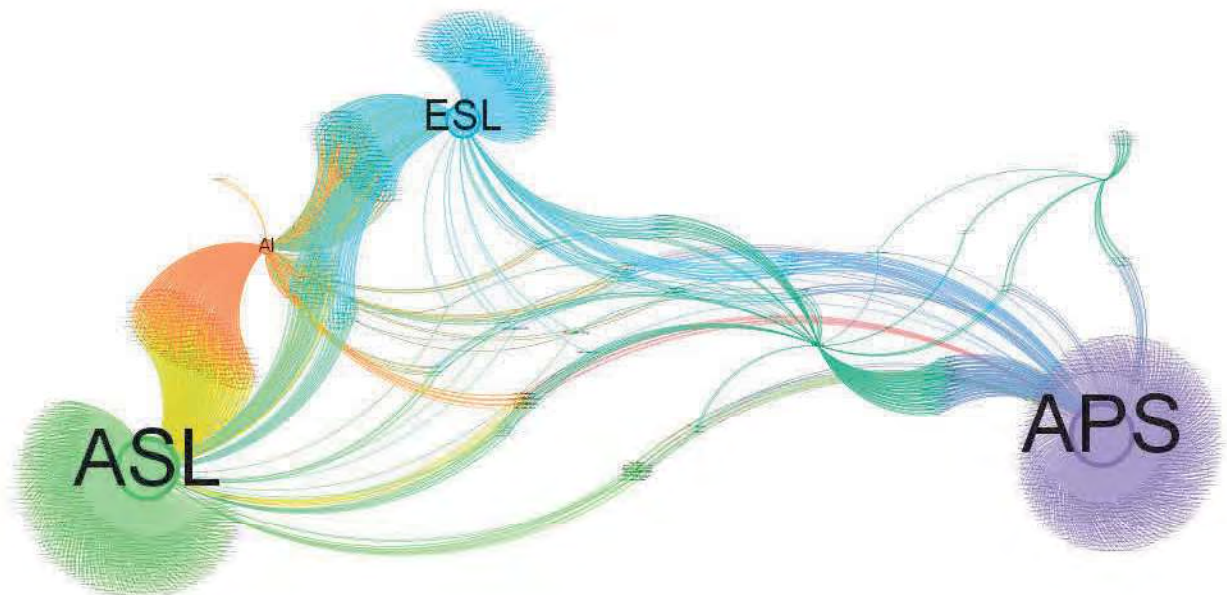
The number of nodes here is 3094 persons plus 6 groups = 3100 (the additional 6 nodes are the CEDA names (CQA, APS, ESL, AI, ASL, and HOD). Gephi will produce a 'bigraph' of the data where relationships are between individual persons and membership organisations.

Code cell 27 - Write the gexf file


```
# Finally we can write a gexf file which will be placed in the root director  
# We can then open the file in Gephi and visualize the network.  
  
nx.write_gexf(G, 'ceda_laidlaw2.gexf')
```

5.29 Visual analysis of the exercise

5.30 The CEDA political network with HOD added



Society	abv.	Dates	Colour
Quaker Committee on the Aborigines and the Thomas Hodgkin MD group	QCA	1832/37 - 1846	Dark green
Aborigines Protection Society	APS	1837 - 1919	Purple
Ethnological Society of London	ESL	1843 - 1871	Blue
Anthropological Society of London	ASL	1863 - 1871	Light Green
Anthropological Institute	AI	1843 - 1871	Orange
Protecting the Empire's Humanity	HOD	2021	dark green

The Hodgkin network and it's relationships (shown above) are much clearer now that the groups WEL and ZOE have been combined into one. (Both of these former groups were identified by Laidlaw in PEH). The graph is best read 'right' to 'left' following the groups shown in dark green. The Quaker Committee on the Aborigines (QCA), active from 1832 - 1837 (or 1846?) appears top right. This group is 'led' by Thomas Hodgkin MD. Below is the APS formed in 1837 and active to 1919 (beyond the entire period studied in the thesis). We know that Thomas Hodgkin MD formed the APS at Ratcliffe Quaker Meeting House in 1837.

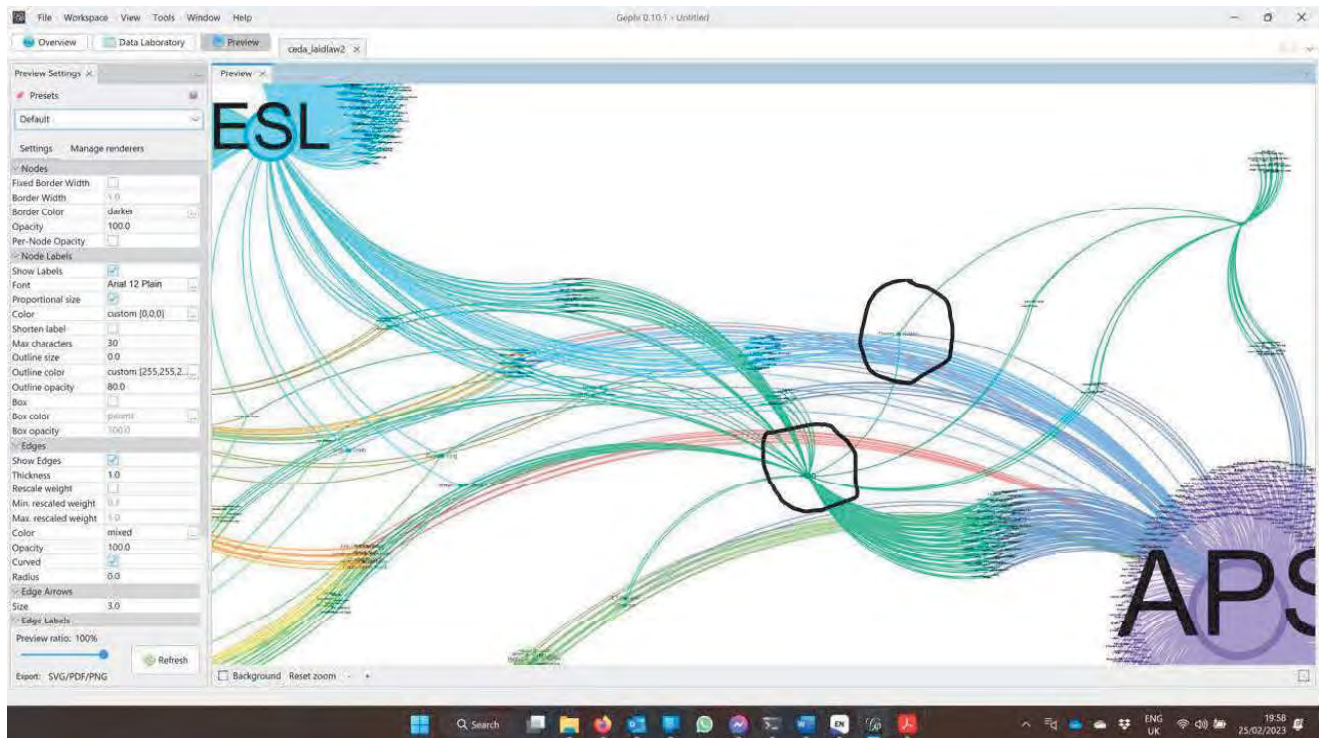
"Dr Hodgkin moved that an auxiliary society be formed in Ratcliff [? Quaker meeting ?] for the purpose of promoting the objects of the Aborigines Protection Society and more especially to collect information from persons recently arrived from abroad and to exert an interest in those who might be going out as colonists or sailors. He observed that though a great amount of valuable information relating to the subjects now under consideration was brought to this part from various parts of the globe very great difficulty had been experienced in collecting it. The formation of the proposed branch society might do much to overcome this difficulty."

(WELLCOME LIBRARY FOR THE HISTORY AND UNDERSTANDING OF MEDICINE
DEPARTMENT OF ARCHIVES AND MANUSCRIPTS. HODGKIN FAMILY PAPERS 1996
PP/HO/D, Thomas Hodgkin MD (1798 1866), General Material on Civilisation and
Colonialization. Aborigines Protection Society General Materials.)

PP/HO/D/D148, Minutes of the First Meeting, (3ff), 1837

Thomas Hodgkin's network (HOD) does not centre closely to the QCA, it lies at the heart of the greater network collected and compiled in the HTTD. It links most strongly with the APS, but also has good connectivity with the ESL and later the ASL. It has poor connectivity to the AI.

5.31 The CEDA political network with HOD added in detail



We can see that Hodgkin sits apart, pulled (By the Force Atlas algorithm) to sit between the QCA, the APS and his own personal network (HOD). He is revealed as highly networked and this indicates a possible key influencer.

5.32 GitHub upload

We can now update GitHub to pass this exercise and all its resources to the dedicated project repo (see step 1). This enables the entire exercise to be both scrutinised by others and replicated elsewhere (When the exercise is completed and audited, and all copyright issue resolved the repo can be made public).

P7 Chapter 6 Case Study 2 589 Quakers

and their family relationships

File Name: jnb_hddt_quaker_tables

6.1 Import resources

```
import csv
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from operator import itemgetter
import networkx as nx
from networkx.algorithms import community
#This part of networkx, for community detection, needs to be imported separately
import nbconvert
import seaborn as sns
plt.rc('figure', figsize=(20, 10))
# 
```

```
-----
ModuleNotFoundError                                Traceback (most recent call last)
Cell In[1], line 3
      1 import csv
      2 import pandas as pd
---->  3 import matplotlib.pyplot as plt
      4 import numpy as np
      5 from operator import itemgetter

ModuleNotFoundError: No module named 'matplotlib'
```

```

quakers = pd.read_csv ('vw_1_quakers.csv')
quakers['birth_year'] = quakers['death_year'].fillna(0).astype(np.int64)
quakers['death_year'] = quakers['death_year'].fillna(0).astype(np.int64)

quaker_relationships = pd.read_csv ('vw_5_person1_person2.csv')
immediate = pd.read_csv ('vw_5_quaker_relationships_3.csv')
close = pd.read_csv ('vw_5_quaker_relationships_2.csv')
distant = pd.read_csv ('vw_5_quaker_relationships_1.csv')

quaker_ceda = pd.read_csv('vw_4_ceda_membership_quakers2.csv')
quaker_aps = pd.read_csv ('vw_4_ceda_membership_quakers_aps2.csv')
quaker_esl = pd.read_csv('vw_4_ceda_membership_quakers_esl2.csv')
quaker_asl = pd.read_csv('vw_4_ceda_membership_quakers_asl2.csv')
quaker_ai = pd.read_csv('vw_4_ceda_membership_quakers_ai2.csv')
quaker_qca = pd.read_csv('vw_4_ceda_membership_quakers_qca2.csv')
quaker_hod = pd.read_csv('vw_4_ceda_membership_quakers_hod2.csv')
quaker_not_hod = pd.read_csv('vw_4_ceda_membership_quakers_not_hod2.csv')

```

Quakers with CEDA memberships but without family relationships are circled in blue

Quakers begin their engagement with the Quaker Committee on the Aborigines. Led by Thomas Hodgkin MD (who can be seen in the centre of the graph) they then become members of the Aborigines Protection Society where they comprise 50% of the members. Some Quakers then go on to be members of the Ethnological Society of London, the Anthropological Society of London, and finally the Anthropological Institute.

6.2 List out all Quakers in the database

```
quakers
```

	Name	birth_year	death_year	data_source_id
0	William Aldam	1890	1890	1
1	S Stafford Allen	1870	1870	1
2	Edward Backhouse	1879	1879	1
3	James (1) Backhouse	1869	1869	1
4	James Bell	1872	1872	1
...
584	Joshua Wilson	0	0	3
585	F Woodhead	0	0	3
586	W Woolston	0	0	3
587	Francis Wright	0	0	3
588	S W Wright	0	0	3

589 rows x 4 columns

6.3 List out all the Quaker family relationships

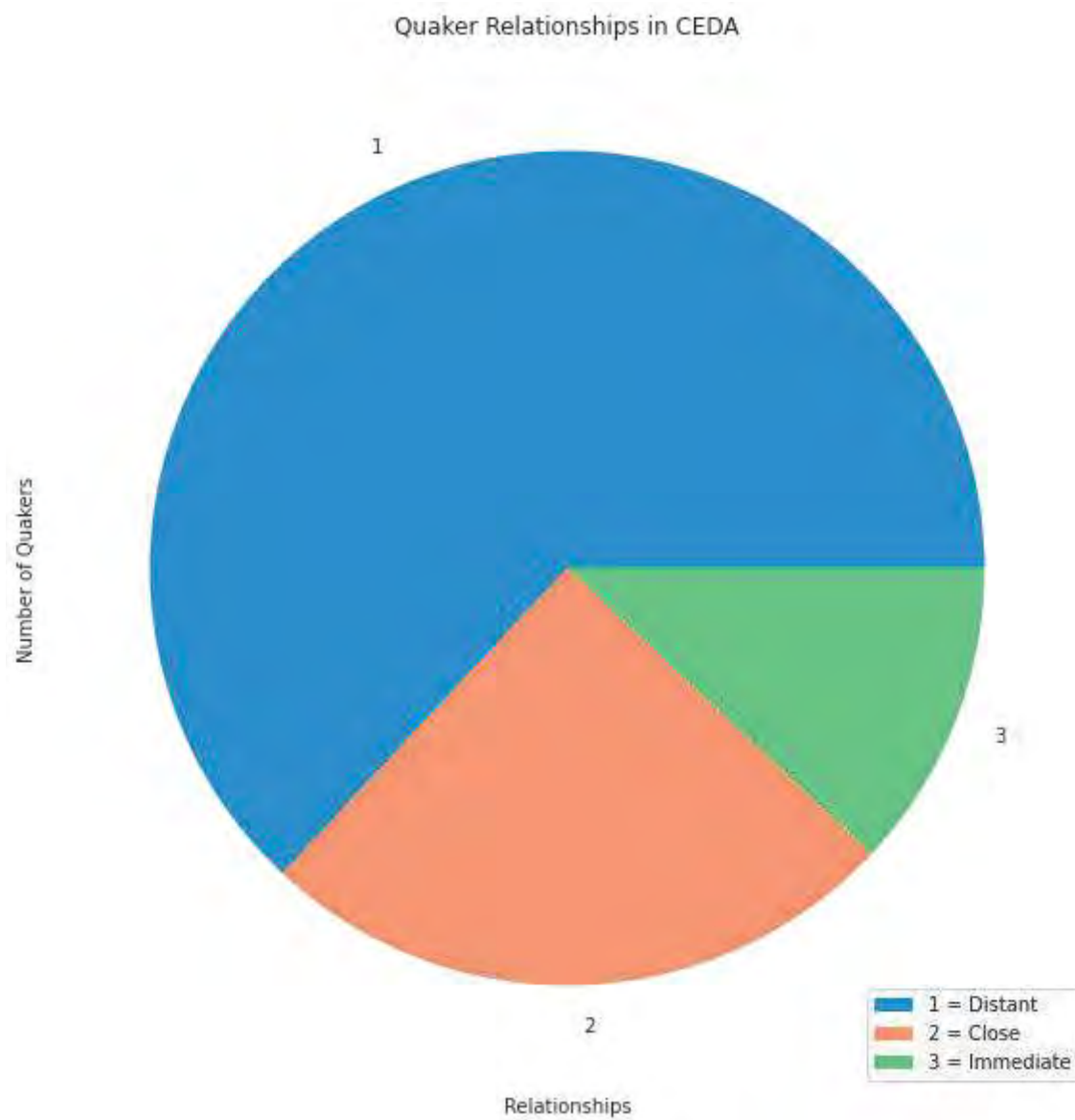
```
quaker_relationships
```

	Source	Target	relationship_type_id
0	William Aldam	x Fox	1
1	William Jun Aldam	x Fox	1
2	Frederick Alexander	R D Alexander	1
3	G W Alexander	R D Alexander	1
4	Henry Alexander	R D Alexander	1
...
2001	Alfred Waterhouse	R Waterhouse	3
2002	Mary Waterhouse	Paul Bevan	3
2003	Lucy Westcombe	Thomas Westcombe	3
2004	Benjamin Wheeler	Samuel Wheeler	3
2005	Charles Wilson	Joshua Wilson	3

2006 rows x 3 columns

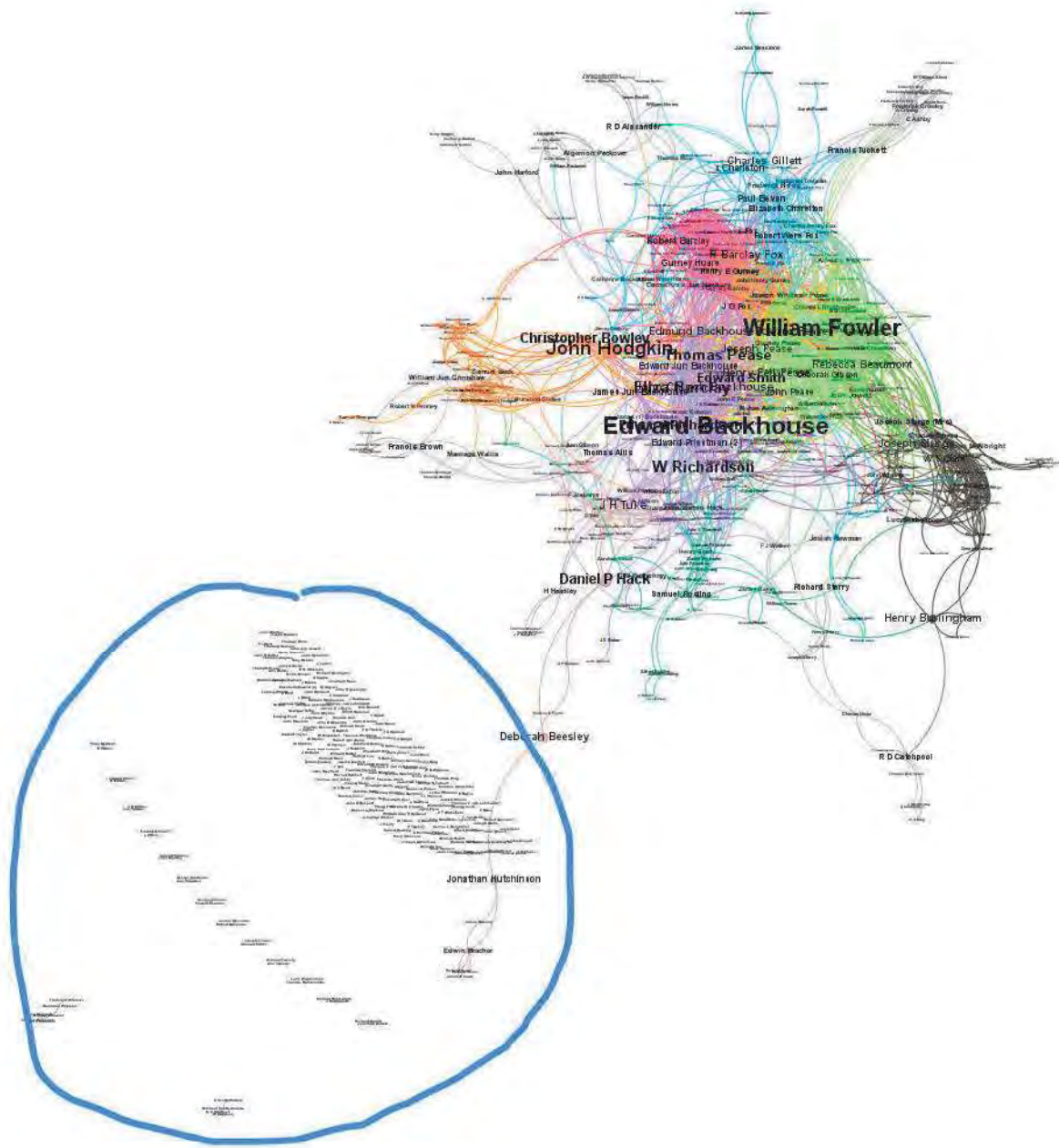
6.4 All Quaker family relationships

```
# quaker_relationships.groupby('relationship_type_id')['Source'].nunique().plot
quaker_relationships.groupby('relationship_type_id')['Source'].count().plot
plt.title ("Quaker Relationships in CEDA")
plt.xlabel ("Relationships")
plt.ylabel ("Number of Quakers")
plt.legend(["1 = Distant", "2 = Close", "3 = Immediate"], loc ="lower right")
plt.show()
```

6.5 Quakers and their family relationship

networks



6.6 Immediate relationships

immediate

	Source	Target	relationship_type_id
0	Arthur Albright	John M Albright	3
1	Arthur Albright	Rachel Albright	3
2	Arthur Albright	William Albright	3
3	Rachel Albright	John M Albright	3
4	Rachel Albright	William Albright	3
...
236	Alfred Waterhouse	R Waterhouse	3
237	Mary Waterhouse	Paul Bevan	3
238	Lucy Westcombe	Thomas Westcombe	3
239	Benjamin Wheeler	Samuel Wheeler	3
240	Charles Wilson	Joshua Wilson	3

241 rows x 3 columns

6.7 Close relationships

close

	Source	Target	relationship_type_id
0	R D Alexander	Christopher Bowley	2
1	R D Alexander	Robert Charleton	2
2	R D Alexander	Frederick H Fox	2
3	R D Alexander	Thomas Maw	2
4	R D Alexander	William Norton	2
...
495	W Whiting	John Whiting	2
496	Isaac Wilson	S Braithwaite	2
497	Isaac Wilson	John Jowett	2
498	Isaac Wilson	John E Wilson	2
499	William Spicer Wood	Daniel Doncaster	2

500 rows × 3 columns

6.8 Distant relationships

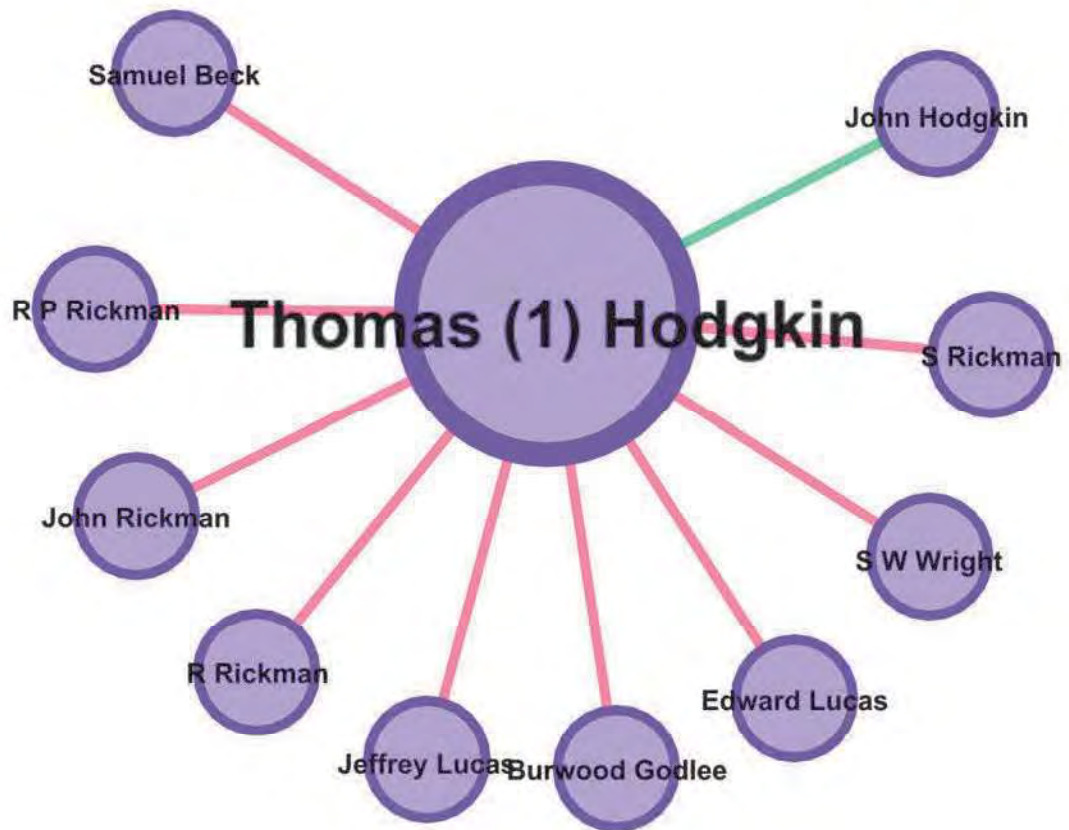
distant

	Source	Target	relationship_type_id
0	William Aldam	x Fox	1
1	William Jun Aldam	x Fox	1
2	Frederick Alexander	R D Alexander	1
3	G W Alexander	R D Alexander	1
4	Henry Alexander	R D Alexander	1
...
1260	William Wilson	Barnard Dickinson	1
1261	William Wilson	Frederick Fryer	1
1262	William Wilson	Benjamin Jowett (2)	1
1263	William Wilson	John Pease	1
1264	William Wilson	Joseph Pease	1

1265 rows × 3 columns

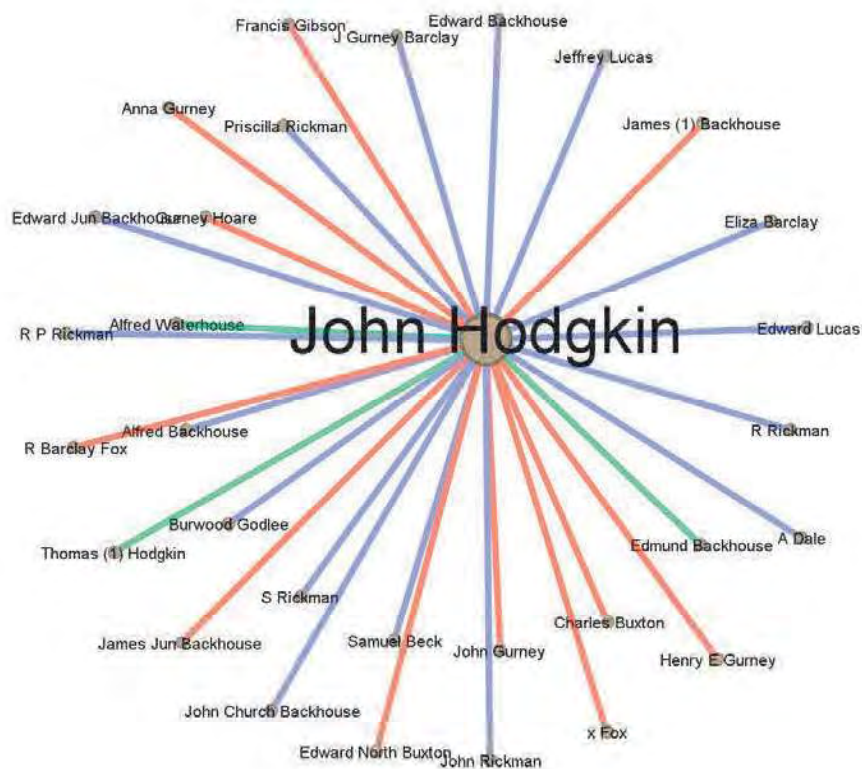
6.9 Significant personal networks - Thomas

Hodgkin MD



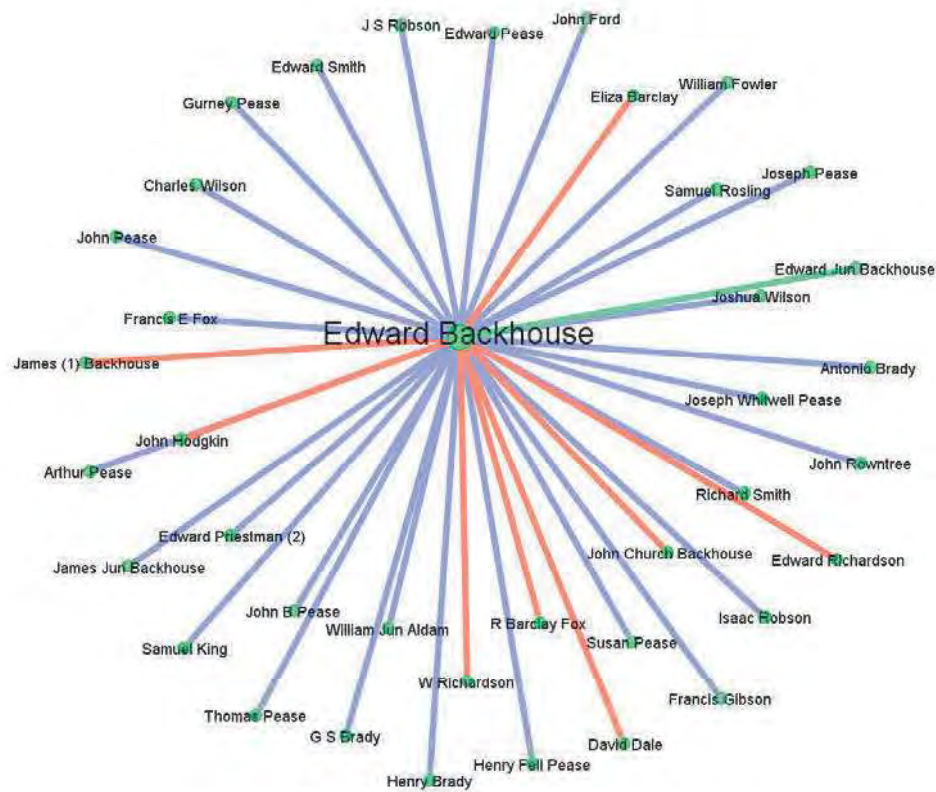
6.10 Significant personal networks - John

Hodgkin

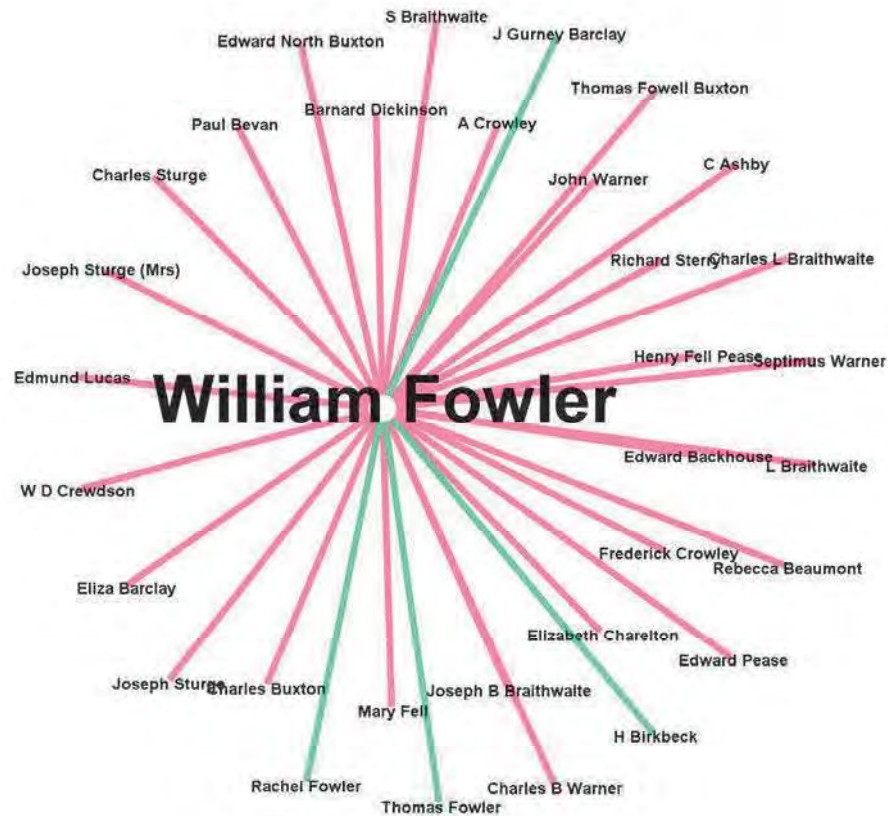


6.11 Significant personal networks - Edward

Backhouse



6.12 Significant personal networks - William



6.13 List out all Quaker members of the CEDA

quaker_ceda

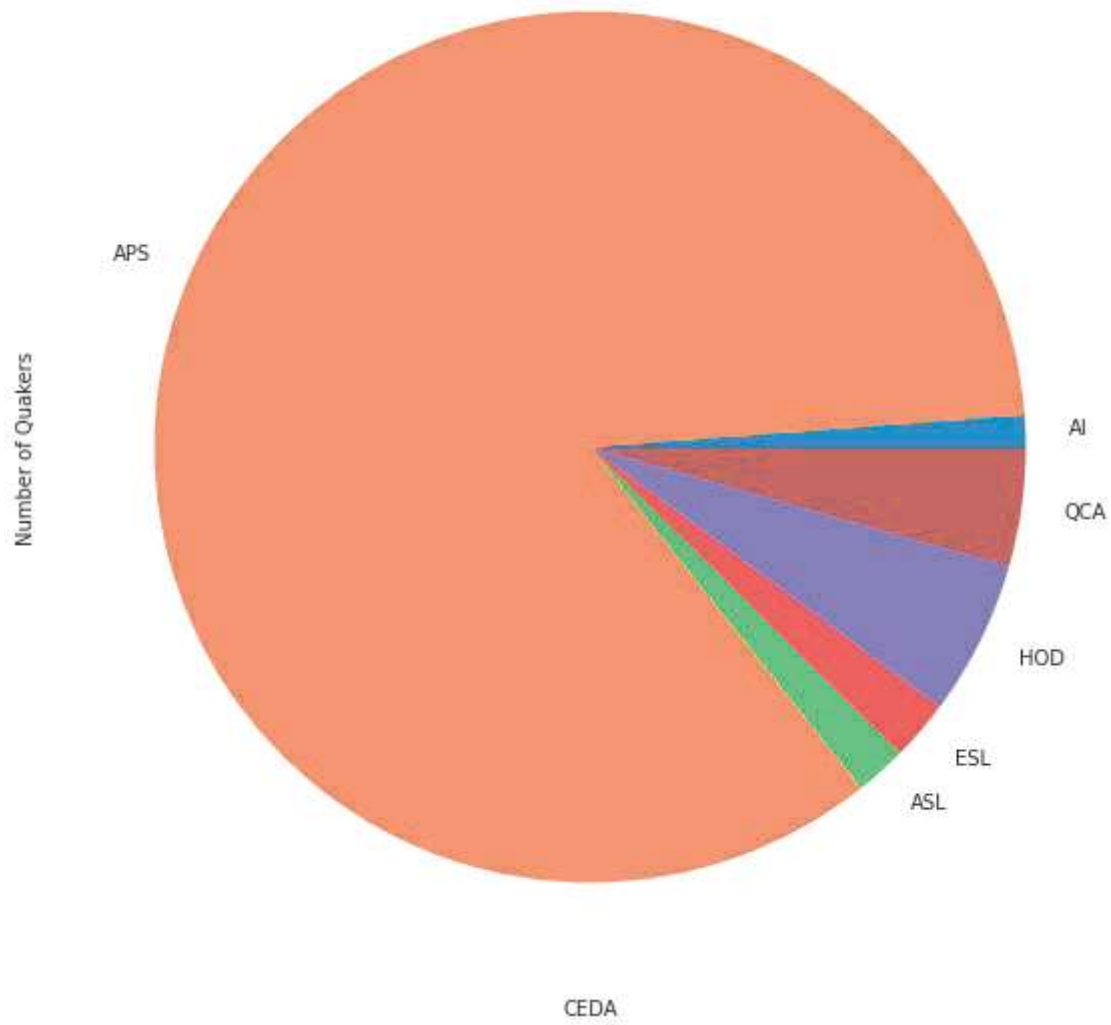
	Name	religion_name	ceda_name	first_year	last_year
0	William Spicer Wood	Quaker	APS	1864.0	1867.0
1	William Spicer Wood	Quaker	ASL	1863.0	1871.0
2	William Spicer Wood	Quaker	AI	1863.0	1871.0
3	William Wilson	Quaker	APS	1838.0	1865.0
4	William Wilson	Quaker	ASL	1865.0	1866.0
...
683	Joshua Wilson	Quaker	APS	1860.0	1860.0
684	F Woodhead	Quaker	APS	1861.0	1862.0
685	W Woolston	Quaker	APS	1861.0	1861.0
686	Francis Wright	Quaker	APS	1838.0	1838.0
687	S W Wright	Quaker	APS	1861.0	1861.0

688 rows x 5 columns

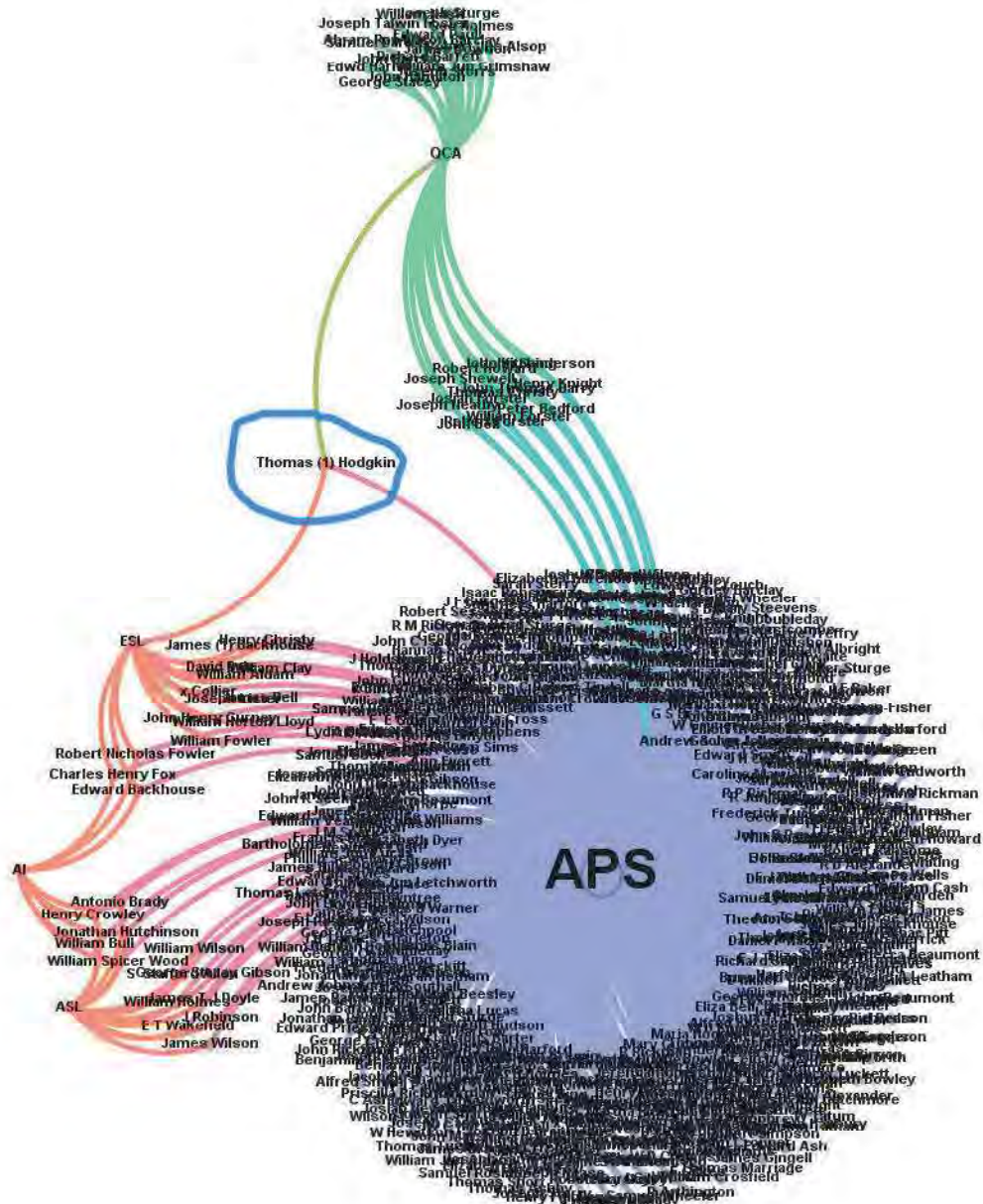
6.14 Pie chart Quaker CEDA memberships

```
quaker_ceda.groupby('ceda_name')['Name'].nunique().plot(kind='pie')
plt.title("Quakers in CEDA")
plt.xlabel("CEDA")
plt.ylabel("Number of Quakers")
plt.show()
```


Quakers in CEDA

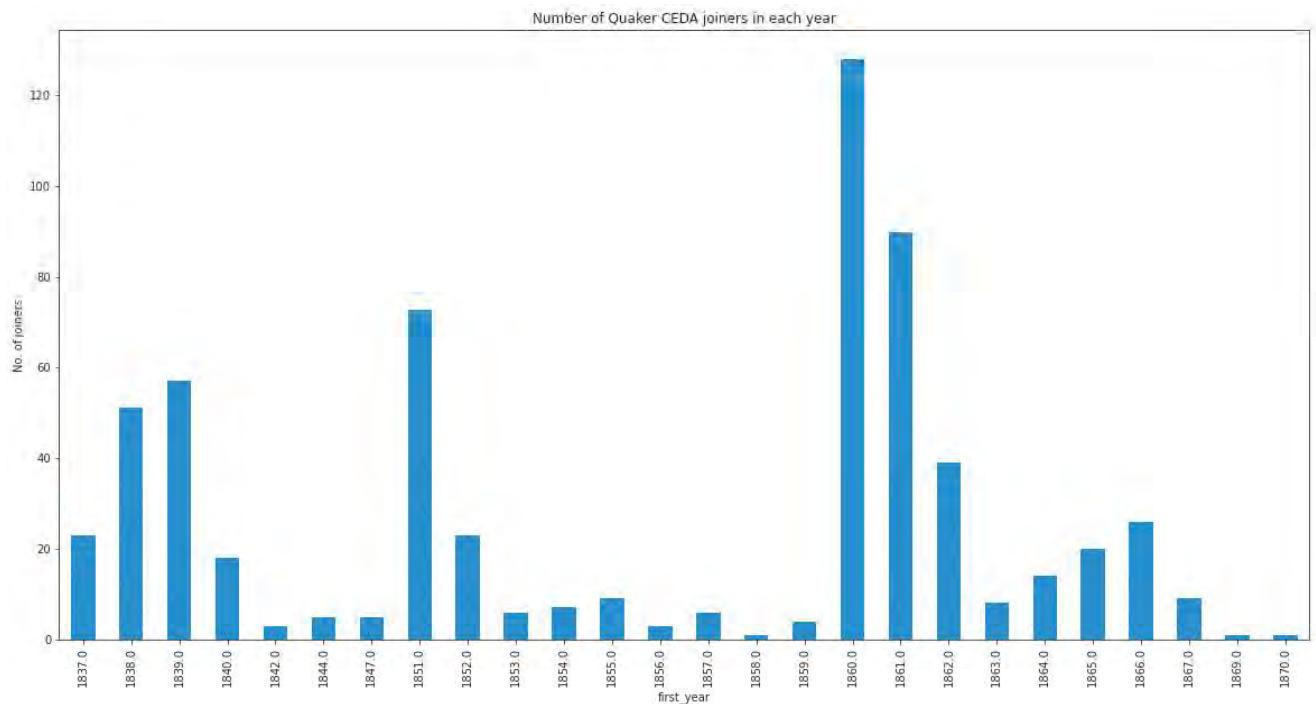


6.15 Quaker CEDA membership networks



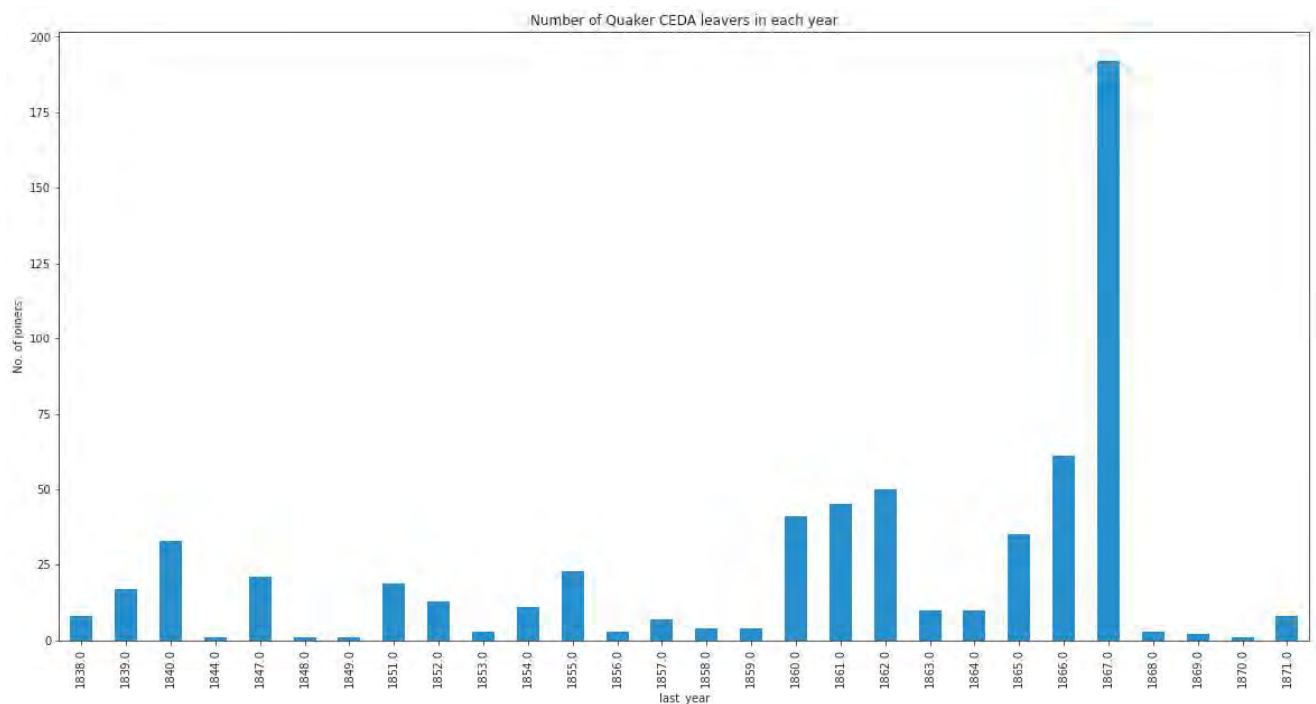
6.16 Quaker joiners of the CEDA by years

```
quaker_ceda.groupby('first_year')['Name'].nunique().plot(kind='bar')
plt.title ("Number of Quaker CEDA joiners in each year")
plt.ylabel ("No. of joiners")
plt.show()
```



6.17 Quaker leavers of the CEDA by years

```
quaker_ceda.groupby('last_year')['Name'].nunique().plot(kind='bar')
plt.title("Number of Quaker CEDA leavers in each year")
plt.ylabel("No. of joiners")
plt.show()
```



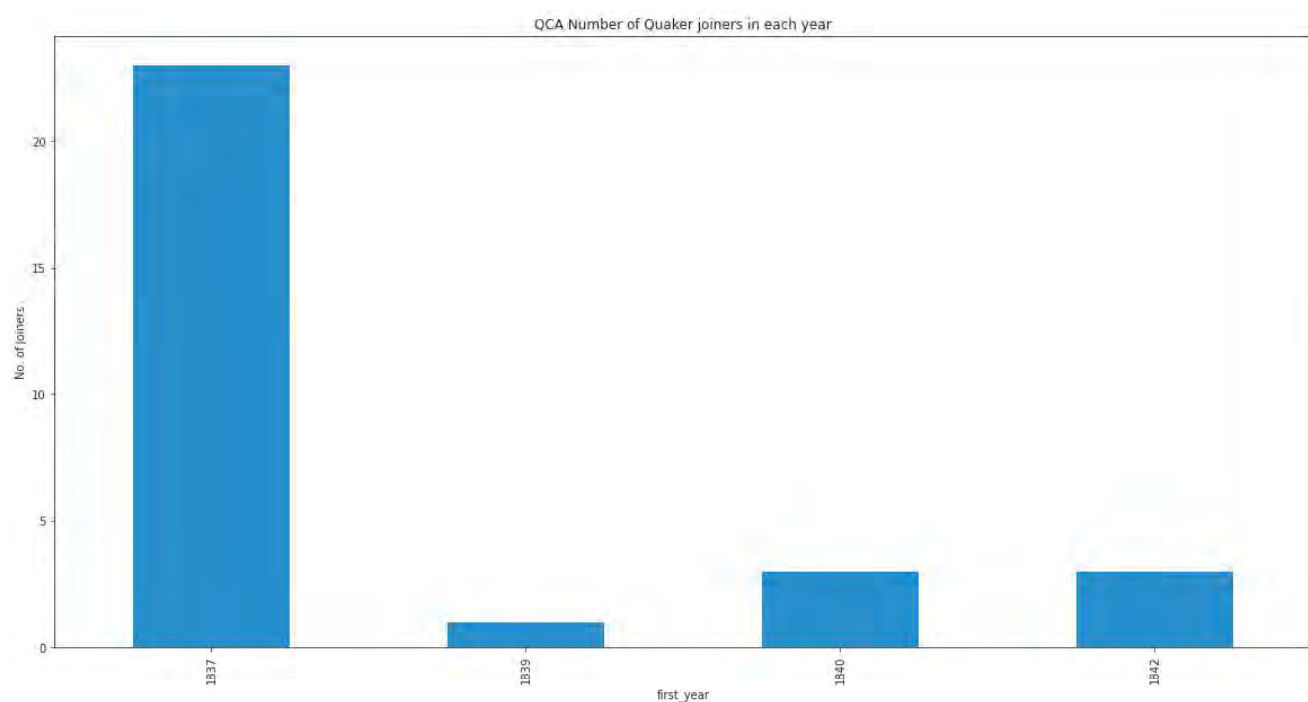
6.18 Quaker members of the QCA

	Name	religion_name	ceda_name	first_year	last_year
0	Thomas (1) Hodgkin	Quaker	QCA	1839	1847
1	James Bowden	Quaker	QCA	1842	1847
2	William Nash	Quaker	QCA	1842	1847
3	Joseph Sturge	Quaker	QCA	1842	1847
4	William Jun Grimshaw	Quaker	QCA	1840	1847
5	Henry Knight	Quaker	QCA	1840	1847
6	Edward Paull	Quaker	QCA	1840	1847
7	Robert Jun Alsop	Quaker	QCA	1837	1847
8	Abram Rawlinson Barclay	Quaker	QCA	1837	1839
9	John Barclay	Quaker	QCA	1837	1839
10	Richard Barrett	Quaker	QCA	1837	1839
11	John Thomas Barry	Quaker	QCA	1837	1847
12	Peter Bedford	Quaker	QCA	1837	1847
13	John Bell	Quaker	QCA	1837	1839
14	Thomas Christy	Quaker	QCA	1837	1839
15	Samuel Darton	Quaker	QCA	1837	1839
16	Josiah Forster	Quaker	QCA	1837	1847
17	Robert Forster	Quaker	QCA	1837	1847
18	William Forster	Quaker	QCA	1837	1847
19	Joseph Talwin Foster	Quaker	QCA	1837	1847
20	John Hamilton	Quaker	QCA	1837	1839
21	Edwd Harris	Quaker	QCA	1837	1847
22	Geo Holmes	Quaker	QCA	1837	1839
23	Robert Howard	Quaker	QCA	1837	1839
24	John Kitching	Quaker	QCA	1837	1839
25	Joseph Neatby	Quaker	QCA	1837	1847
26	John Sanderson	Quaker	QCA	1837	1847
27	Joseph Shewell	Quaker	QCA	1837	1839
28	George Stacey	Quaker	QCA	1837	1847

	Name	religion_name	ceda_name	first_year	last_year
29	Joseph Storrs	Quaker	QCA	1837	1847

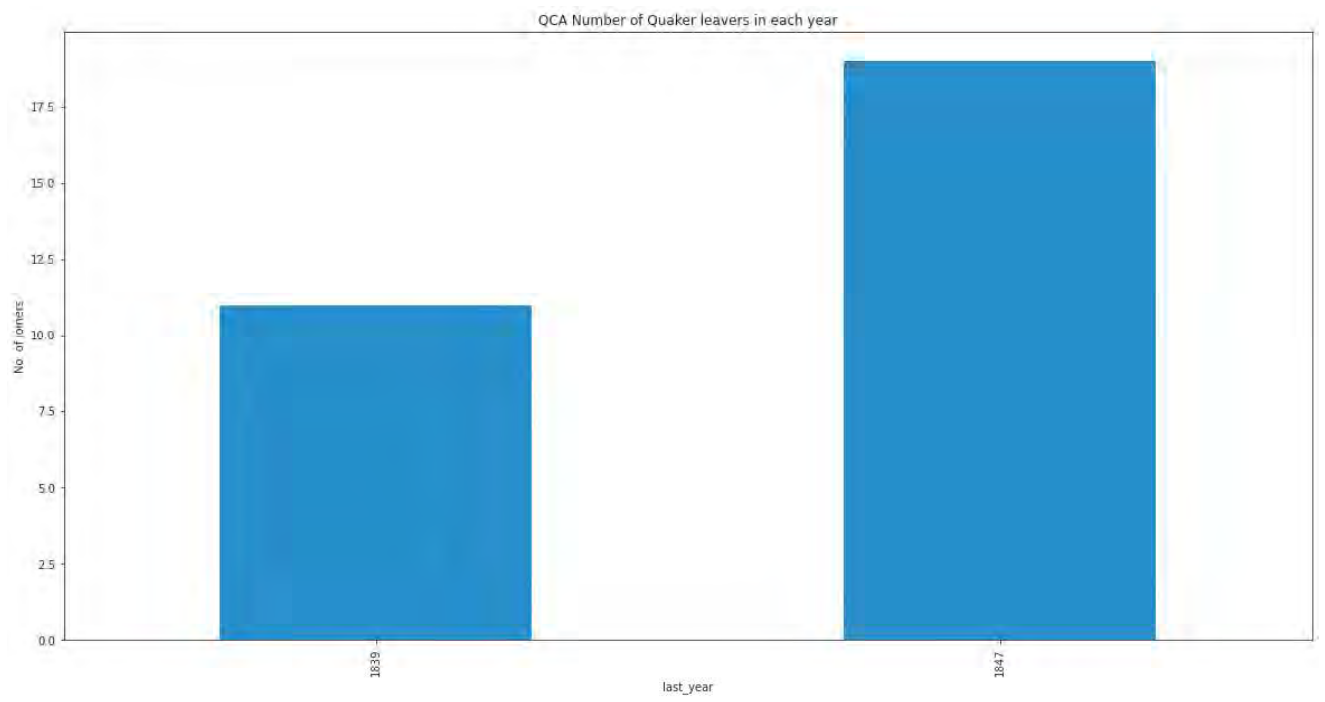
6.19 Quaker joiners of the QCA in years

```
quaker_qca.groupby('first_year')['Name'].nunique().plot(kind='bar')
plt.title ("QCA Number of Quaker joiners in each year")
plt.ylabel ("No. of joiners")
plt.show()
```

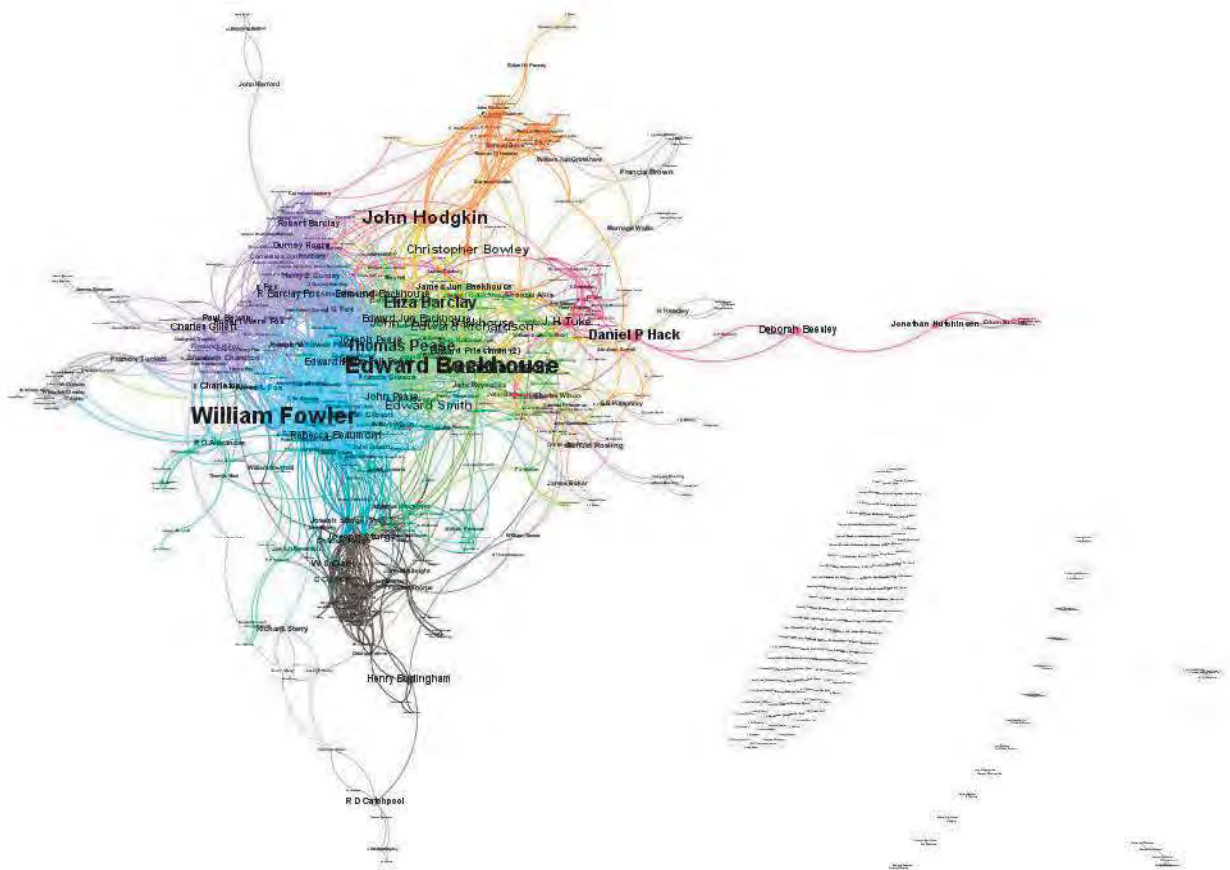


6.20 Quaker leavers of the QCA in years

```
quaker_qca.groupby('last_year')['Name'].nunique().plot(kind='bar')
plt.title ("QCA Number of Quaker leavers in each year")
plt.ylabel ("No. of joiners")
plt.show()
```



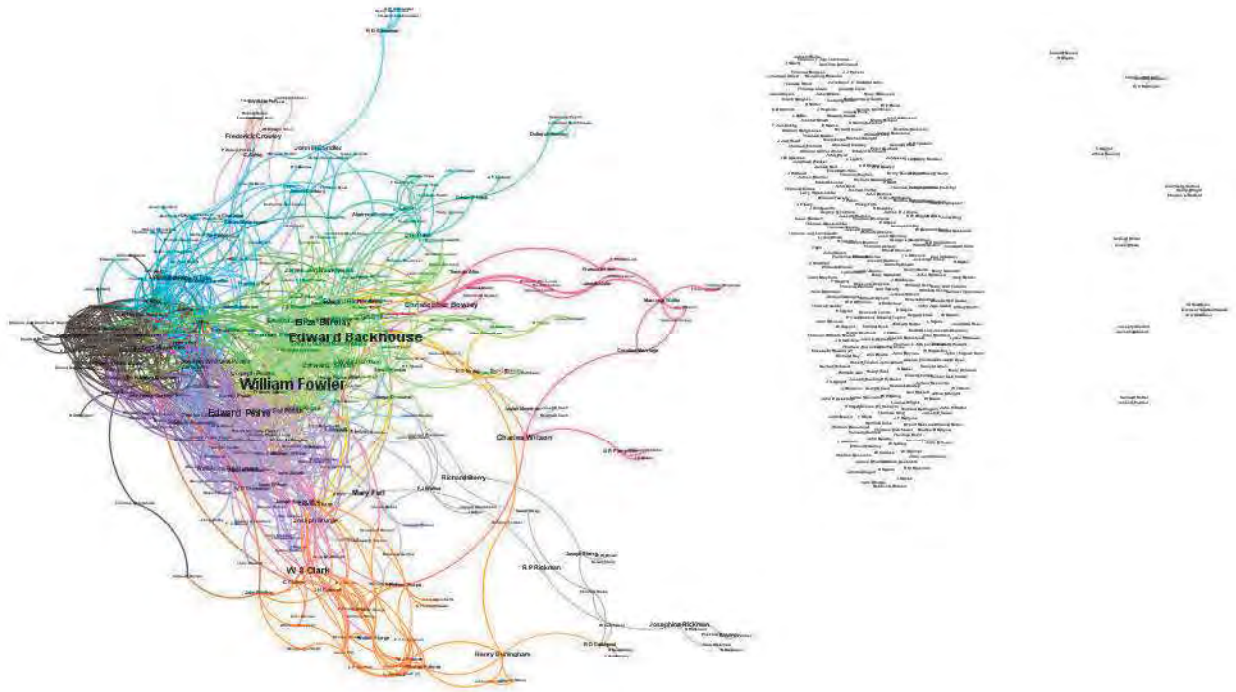
6.21 Show the Quaker members of the APS



Note Quakers make up roughly half of the members over all years

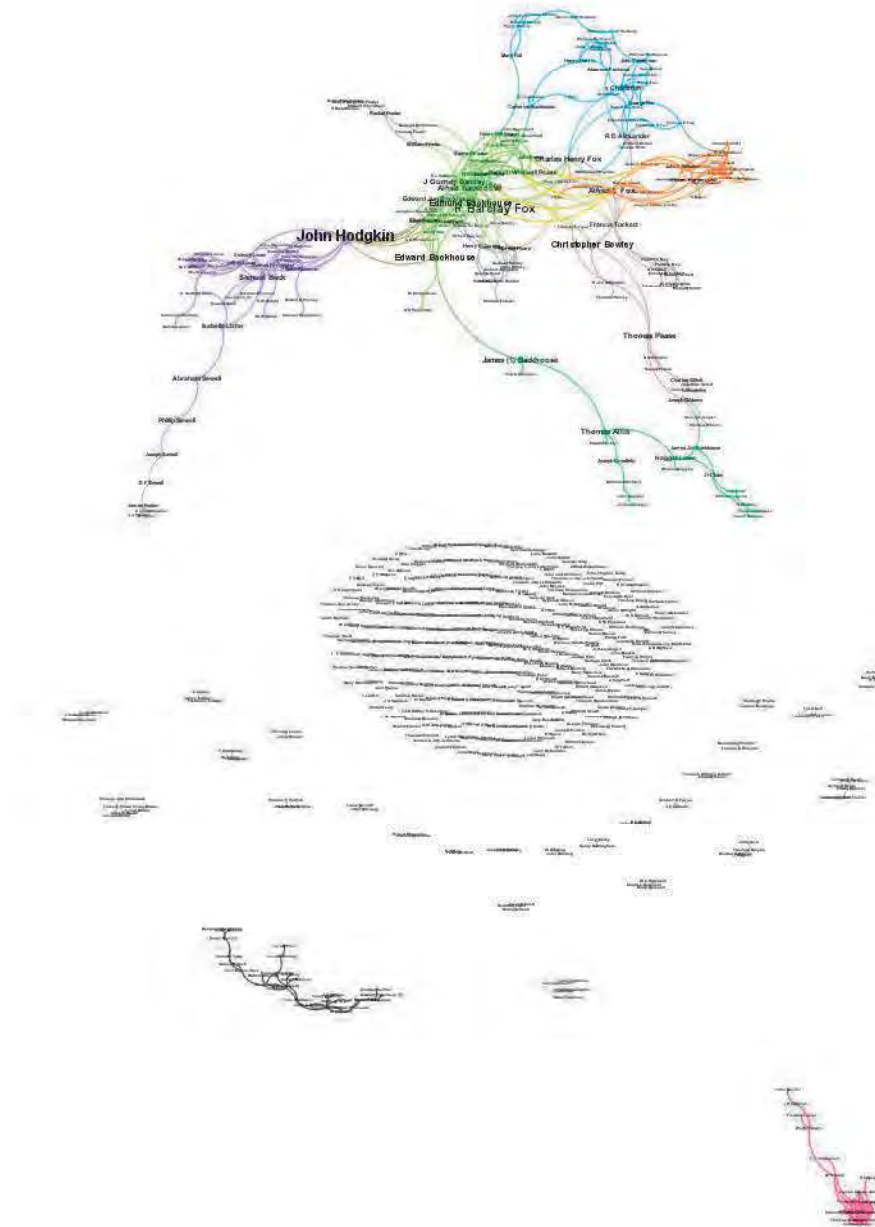
6.22 Quaker members of the APS – distant

relationships



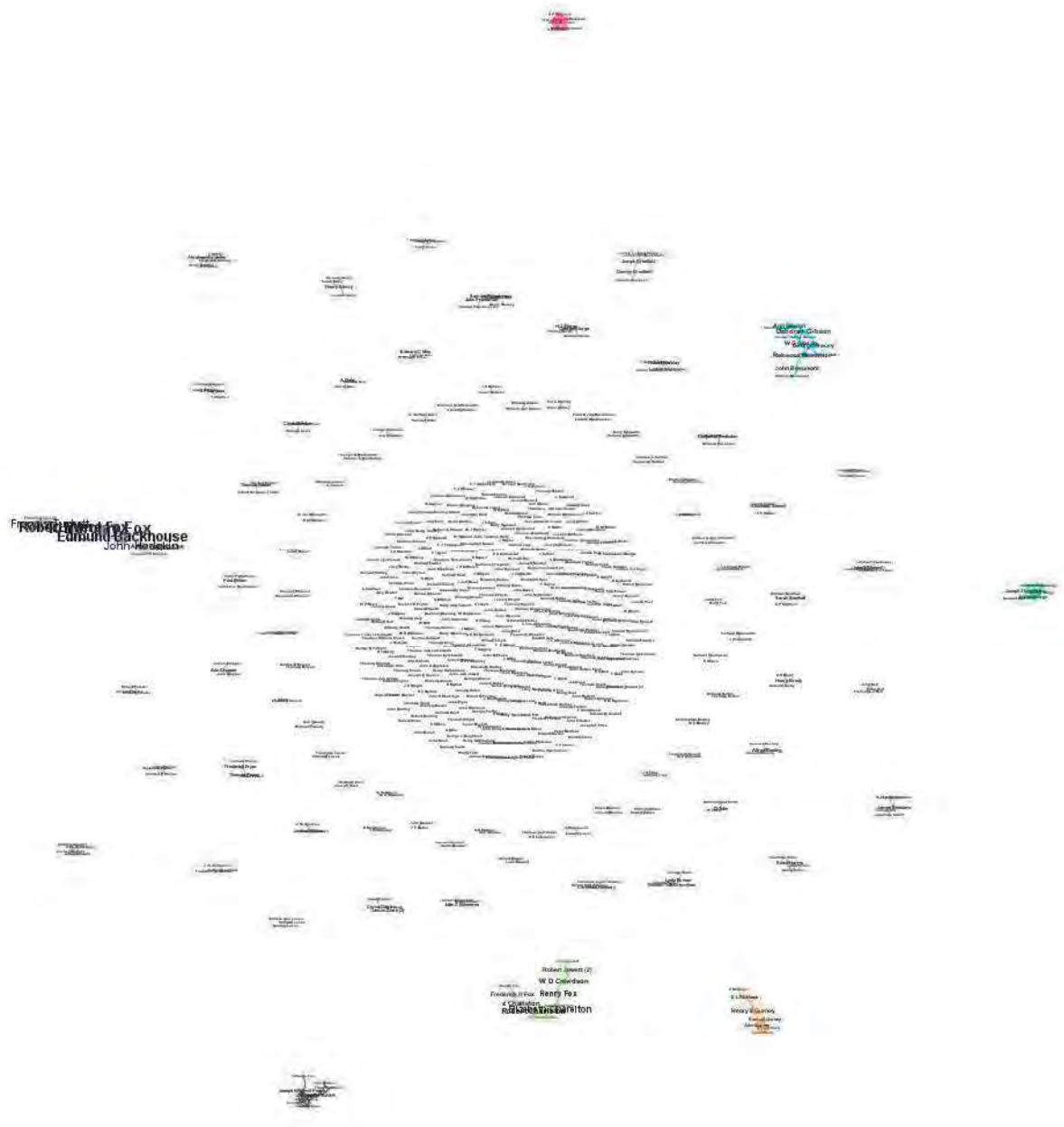
6.23 Quaker members of the APS - close

relationships



6.24 Quaker members of the APS -

immediate relationships

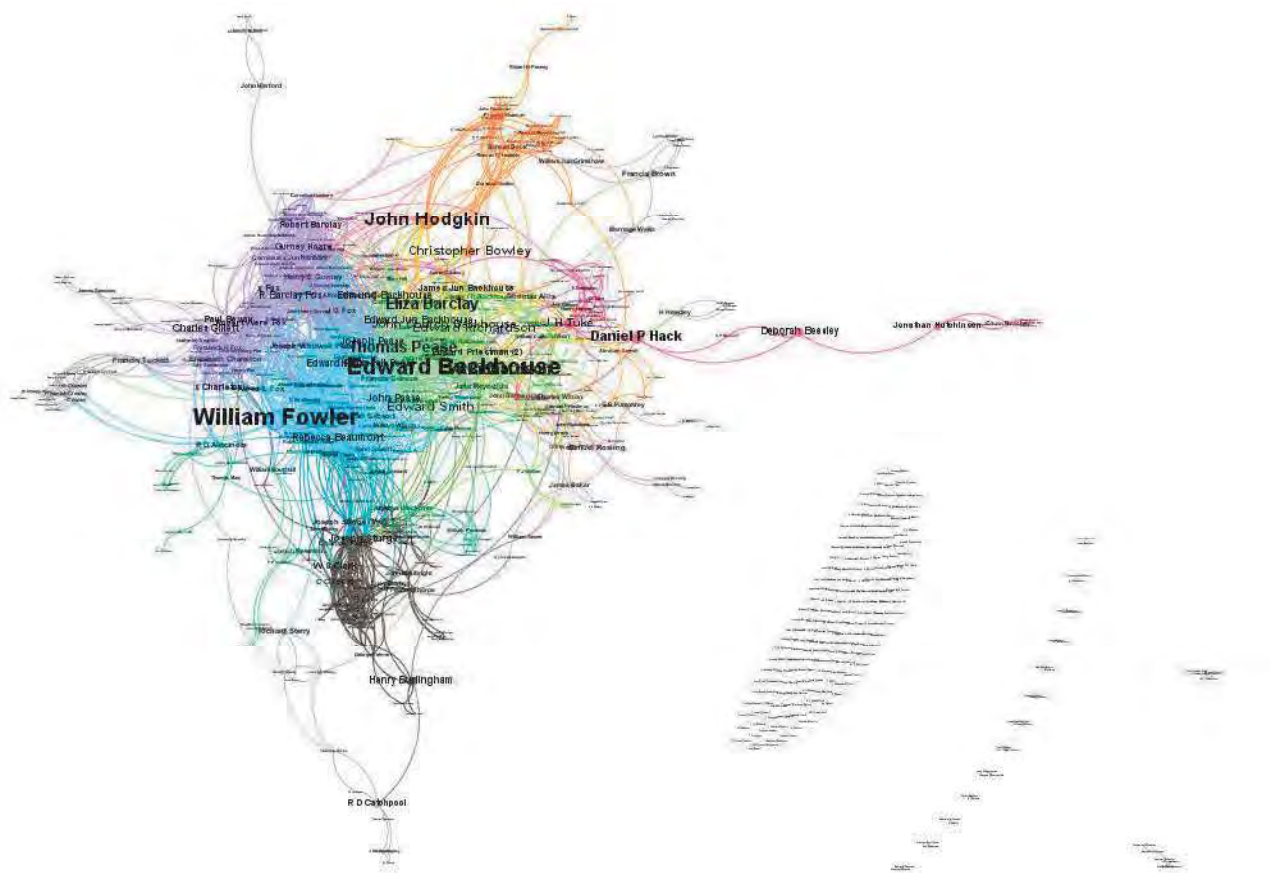


6.25 Show quaker members of the APS

quaker_aps

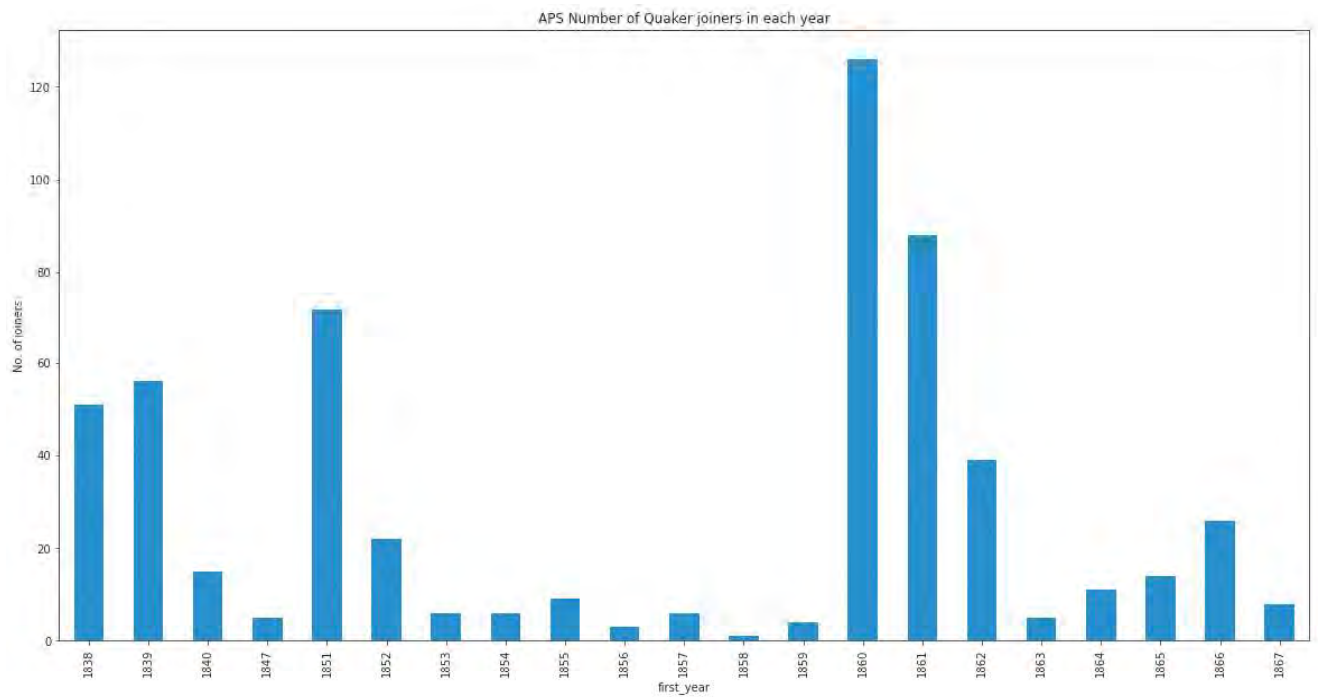
	Name	religion_name	ceda_name	first_year	last_year
0	William Spicer Wood	Quaker	APS	1864	1867
1	William Wilson	Quaker	APS	1838	1865
2	James Wilson	Quaker	APS	1862	1867
3	E T Wakefield	Quaker	APS	1853	1864
4	John Ross	Quaker	APS	1839	1852
...
568	Joshua Wilson	Quaker	APS	1860	1860
569	F Woodhead	Quaker	APS	1861	1862
570	W Woolston	Quaker	APS	1861	1861
571	Francis Wright	Quaker	APS	1838	1838
572	S W Wright	Quaker	APS	1861	1861

573 rows x 5 columns



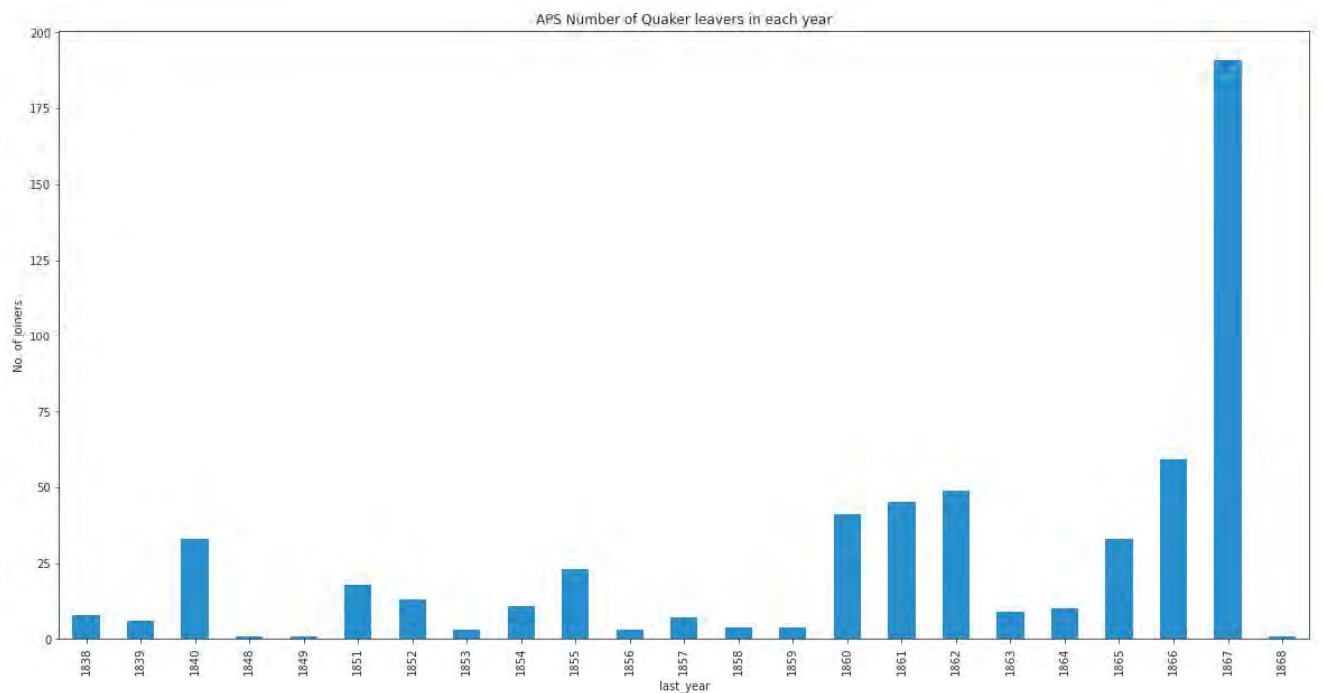
6.26 Pie chart Quaker joiners of the APS

```
quaker_aps.groupby('first_year')['Name'].nunique().plot(kind='bar')
plt.title ("APS Number of Quaker joiners in each year")
plt.ylabel ("No. of joiners")
plt.show()
```



6.27 Pie chart Quaker leavers of the APS

```
quaker_aps.groupby('last_year')['Name'].nunique().plot(kind='bar')
plt.title("APS Number of Quaker leavers in each year")
plt.ylabel("No. of joiners")
plt.show()
```

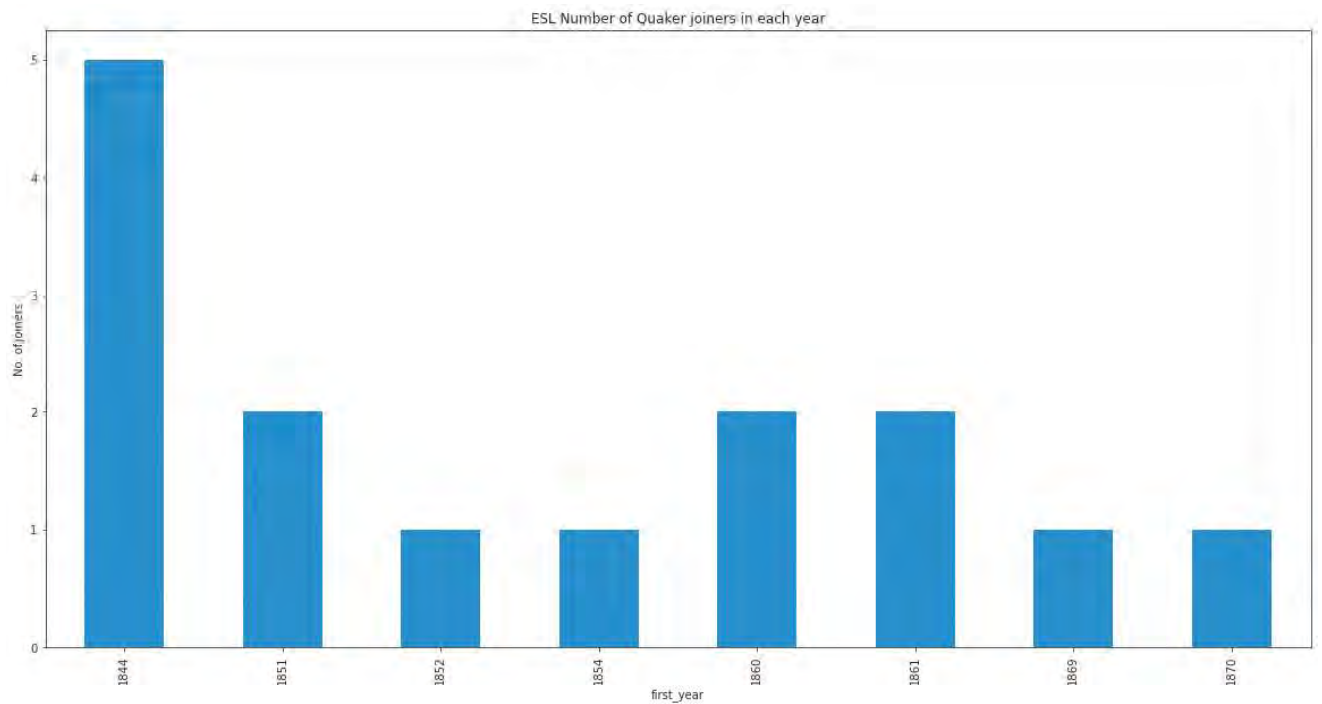


6.28 Show Quaker members of the ESL

	Name	religion_name	ceda_name	first_year	last_year
0	William Horton Lloyd	Quaker	ESL	1844	1847
1	Joseph Lister	Quaker	ESL	1844	1847
2	Thomas (1) Hodgkin	Quaker	ESL	1844	1862
3	John Henry Gurney	Quaker	ESL	1860	1867
4	Charles Henry Fox	Quaker	ESL	1861	1871
5	William Fowler	Quaker	ESL	1851	1851
6	Robert Nicholas Fowler	Quaker	ESL	1851	1871
7	David Dale	Quaker	ESL	1860	1863
8	x Collier	Quaker	ESL	1844	1844
9	William Clay	Quaker	ESL	1861	1868
10	Henry Christy	Quaker	ESL	1854	1865
11	James Bell	Quaker	ESL	1852	1862
12	James (1) Backhouse	Quaker	ESL	1869	1869
13	Edward Backhouse	Quaker	ESL	1870	1871
14	William Aldam	Quaker	ESL	1844	1848

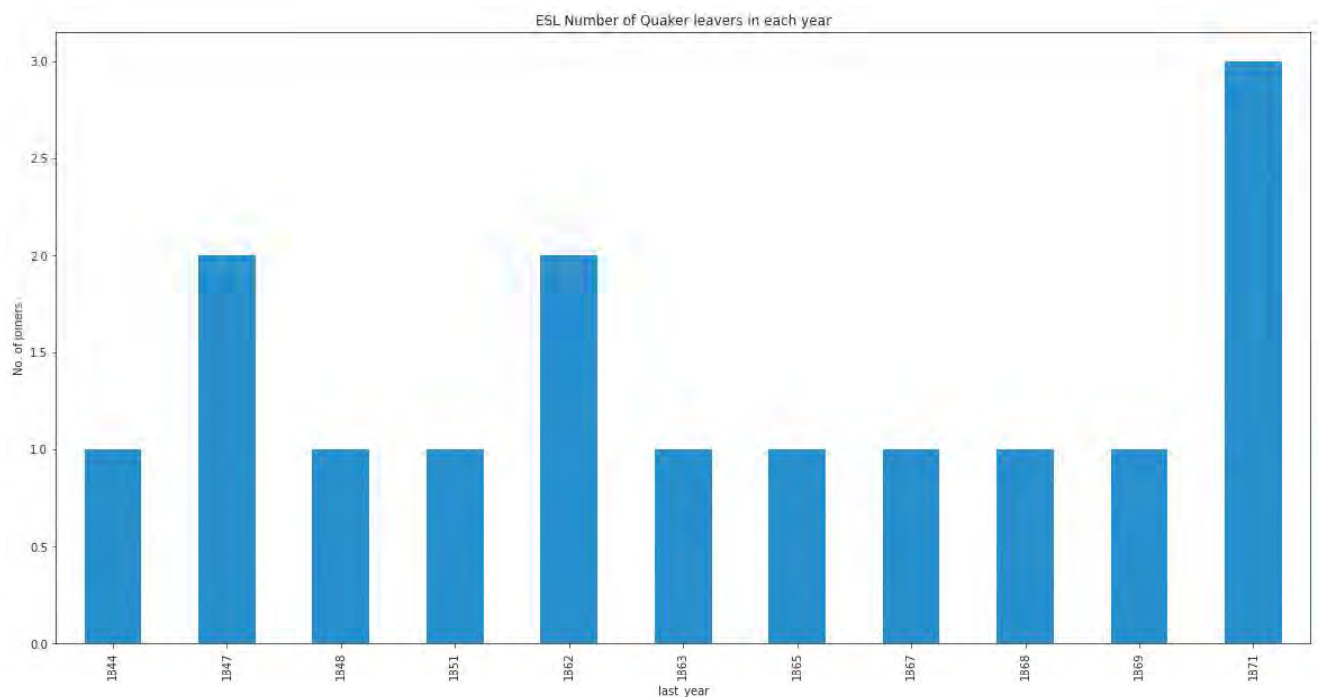
6.29 Show Quaker joiners of the ESL

```
quaker_esl.groupby('first_year')['Name'].nunique().plot(kind='bar')
plt.title("ESL Number of Quaker joiners in each year")
plt.ylabel("No. of joiners")
plt.show()
```



6.30 Show Quaker leavers of the ESL

```
quaker_esl.groupby('last_year')['Name'].nunique().plot(kind='bar')
plt.title("ESL Number of Quaker leavers in each year")
plt.ylabel("No. of joiners")
plt.show()
```

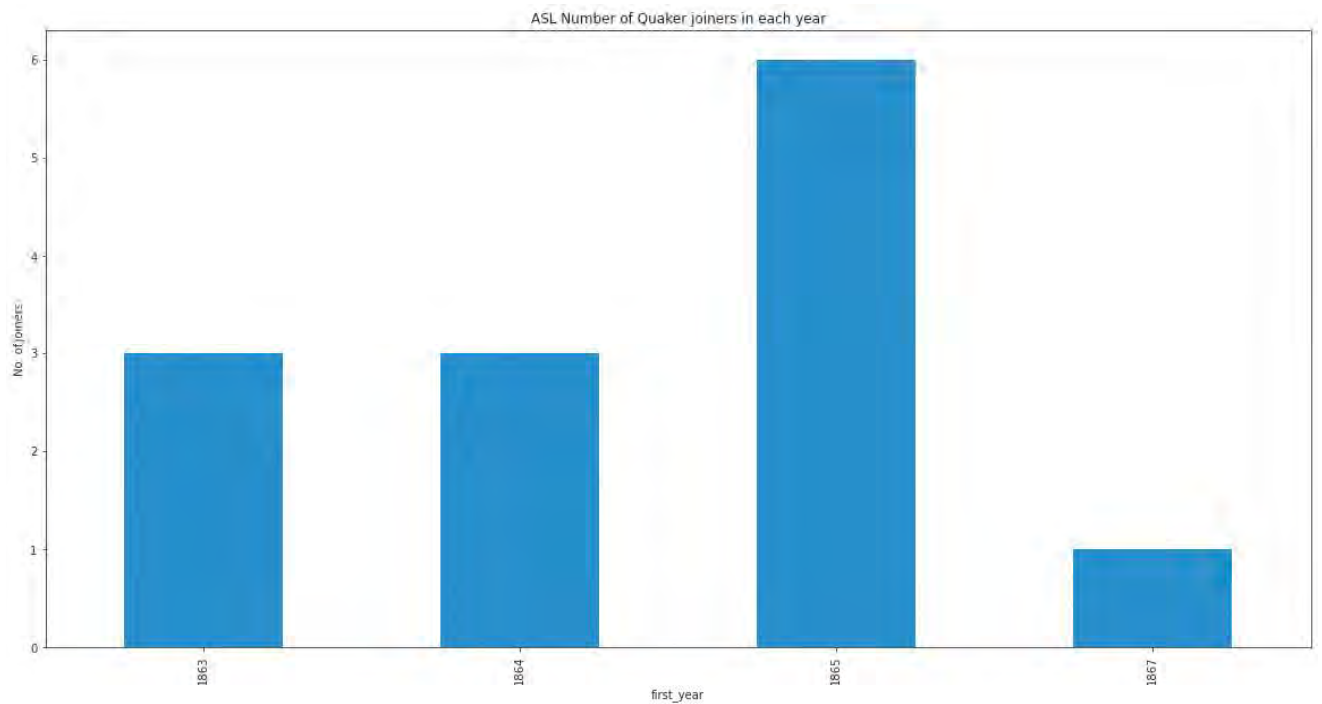


6.31 Show Quaker members of the ASL

	Name	religion_name	ceda_name	first_year	last_year
0	William Spicer Wood	Quaker	ASL	1863	1871
1	William Wilson	Quaker	ASL	1865	1866
2	James Wilson	Quaker	ASL	1865	1865
3	E T Wakefield	Quaker	ASL	1865	1868
4	J Robinson	Quaker	ASL	1865	1865
5	Jonathan Hutchinson	Quaker	ASL	1863	1871
6	William Holmes	Quaker	ASL	1865	1869
7	George Stacey Gibson	Quaker	ASL	1864	1866
8	James T J Doyle	Quaker	ASL	1865	1868
9	Henry Crowley	Quaker	ASL	1864	1871
10	William Bull	Quaker	ASL	1867	1871
11	Antonio Brady	Quaker	ASL	1864	1871
12	S Stafford Allen	Quaker	ASL	1863	1870

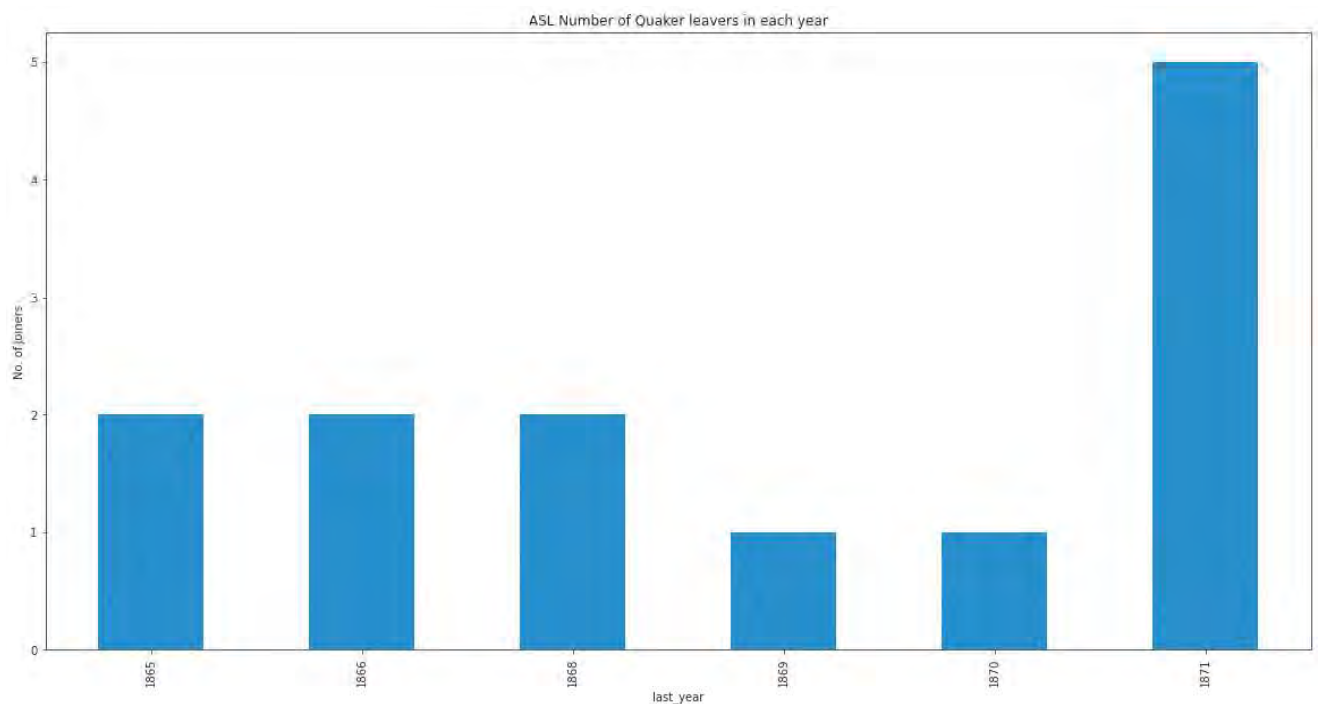
6.32 Show quaker joiners of the ASL

```
quaker_asl.groupby('first_year')['Name'].nunique().plot(kind='bar')
plt.title("ASL Number of Quaker joiners in each year")
plt.ylabel("No. of joiners")
plt.show()
```



6.33 Show quaker leavers of the ASL

```
quaker_asl.groupby('last_year')['Name'].nunique().plot(kind='bar')
plt.title("ASL Number of Quaker leavers in each year")
plt.ylabel("No. of joiners")
plt.show()
```

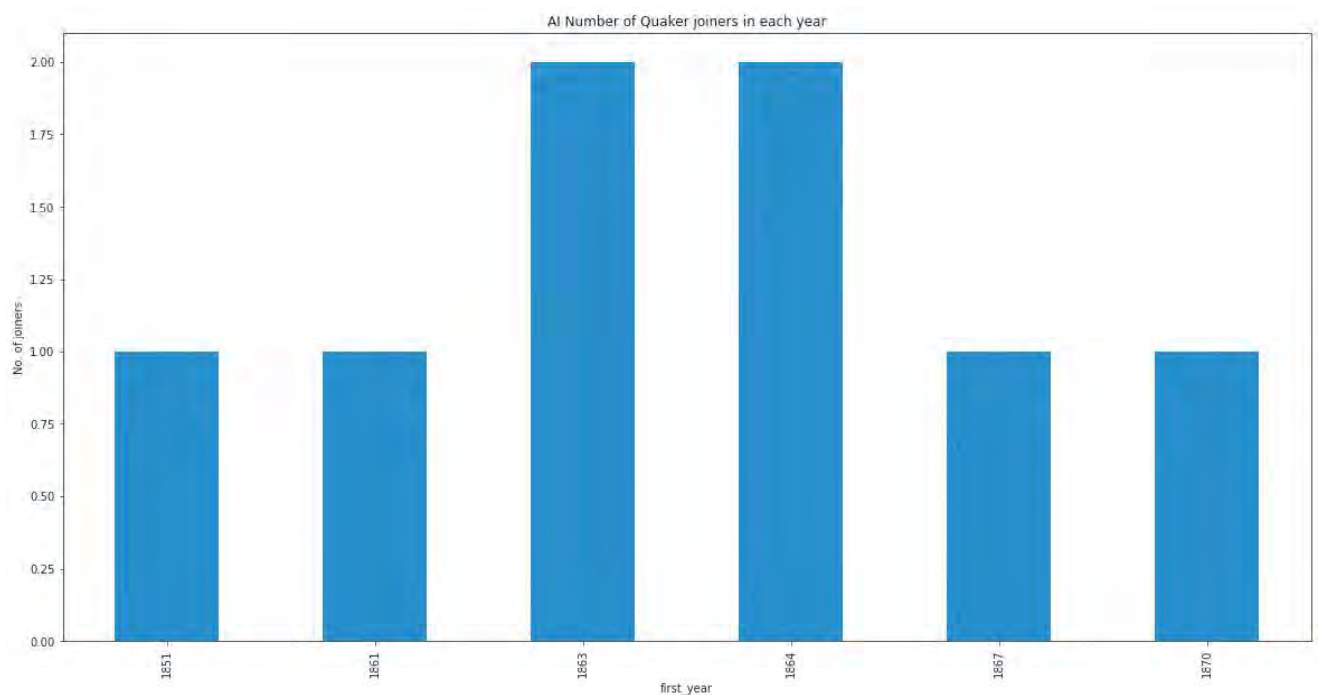


6.34 Show Quaker members of the AI

	Name	religion_name	ceda_name	first_year	last_year
0	William Spicer Wood	Quaker	AI	1863	1871
1	Jonathan Hutchinson	Quaker	AI	1863	1871
2	Charles Henry Fox	Quaker	AI	1861	1871
3	Robert Nicholas Fowler	Quaker	AI	1851	1871
4	Henry Crowley	Quaker	AI	1864	1871
5	William Bull	Quaker	AI	1867	1871
6	Antonio Brady	Quaker	AI	1864	1871
7	Edward Backhouse	Quaker	AI	1870	1871

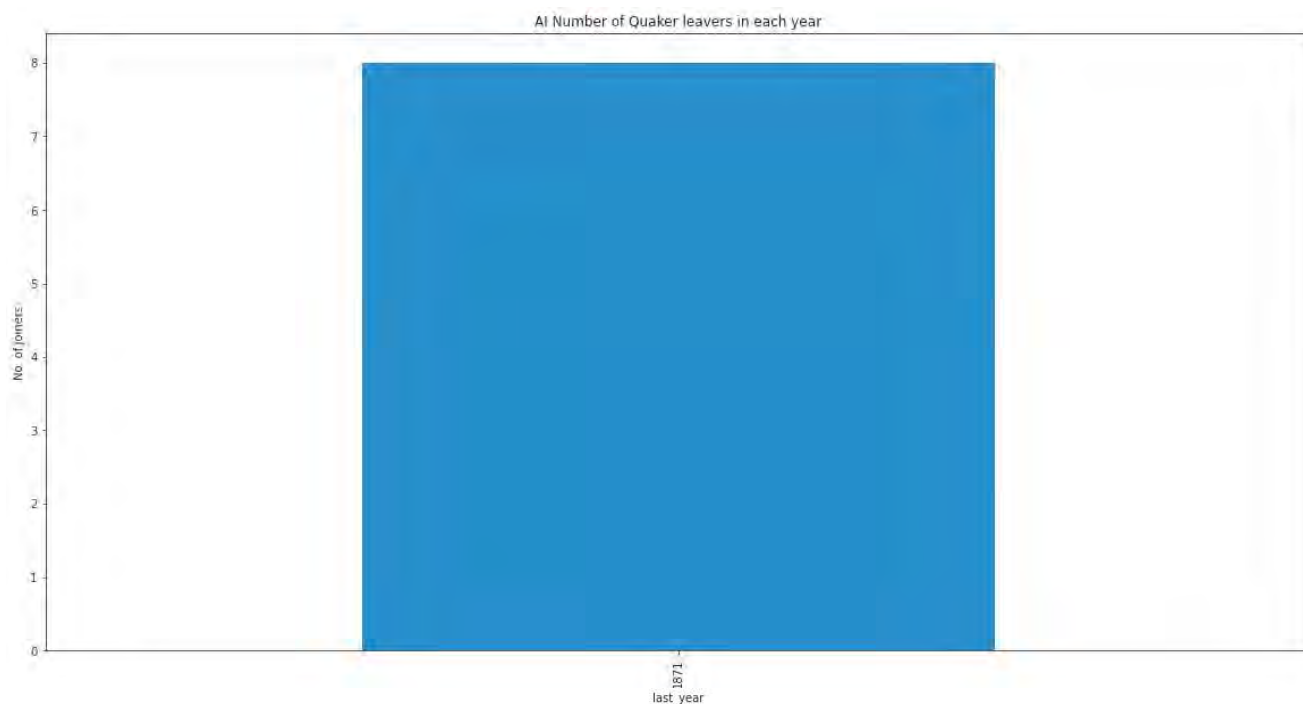
6.35 Show Quaker joiners of the AI

```
quaker_ai.groupby('first_year')['Name'].nunique().plot(kind='bar')
plt.title("AI Number of Quaker joiners in each year")
plt.ylabel("No. of joiners")
plt.show()
```



6.36 Show Quaker leavers of the AI

```
quaker_ai.groupby('last_year')['Name'].nunique().plot(kind='bar')
plt.title ("AI Number of Quaker leavers in each year")
plt.ylabel ("No. of joiners")
plt.show()
```



6.37 Quaker CEDA members of Case Study Three Hodgkin's network

Note There are no 'first' and 'last' year data for this dummy CEDA

```
quaker_hod
```

	Name	birth_year	death_year	religion_name	ceda_name
0	E T Wakefield	NaN	NaN	Quaker	HOD
1	John Ross	NaN	NaN	Quaker	HOD
2	J Robinson	NaN	NaN	Quaker	HOD
3	Joseph Lister	1827.0	1912.0	Quaker	HOD
4	Thomas (1) Hodgkin	1798.0	1866.0	Quaker	HOD
5	Thomas (1) Hodgkin	1798.0	1866.0	Quaker	HOD
6	Robert Nicholas Fowler	1828.0	1891.0	Quaker	HOD
7	Henry Christy	1810.0	1865.0	Quaker	HOD
8	James (1) Backhouse	1794.0	1869.0	Quaker	HOD
9	James (1) Backhouse	1794.0	1869.0	Quaker	HOD
10	Jonathan Backhouse	NaN	NaN	Quaker	HOD
11	W G Gibson	NaN	NaN	Quaker	HOD
12	x Hadfield	NaN	NaN	Quaker	HOD
13	Thomas Harvey	NaN	NaN	Quaker	HOD
14	Thomas Hughes	NaN	NaN	Quaker	HOD
15	S Rickman	NaN	NaN	Quaker	HOD
16	Richard Smith	NaN	NaN	Quaker	HOD
17	W Thompson	NaN	NaN	Quaker	HOD
18	Frederick Tuckett	NaN	NaN	Quaker	HOD
19	Frederick Tuckett	NaN	NaN	Quaker	HOD
20	Joseph Sturge	NaN	NaN	Quaker	HOD
21	Robert Jun Alsop	NaN	NaN	Quaker	HOD
22	Peter Bedford	NaN	NaN	Quaker	HOD
23	Josiah Forster	NaN	NaN	Quaker	HOD
24	Josiah Forster	NaN	NaN	Quaker	HOD
25	William Forster	NaN	NaN	Quaker	HOD
26	Robert Howard	NaN	NaN	Quaker	HOD
27	Robert Howard	NaN	NaN	Quaker	HOD
28	R D Alexander	NaN	NaN	Quaker	HOD

	Name	birth_year	death_year	religion_name	ceda_name
29	J Gurney Barclay	NaN	NaN	Quaker	HOD
30	Joseph B Braithwaite	NaN	NaN	Quaker	HOD
31	Edward Carroll	NaN	NaN	Quaker	HOD
32	Elliott Cresson	NaN	NaN	Quaker	HOD
33	Elliott Cresson	NaN	NaN	Quaker	HOD
34	James Cropper	NaN	NaN	Quaker	HOD
35	Burwood Godlee	NaN	NaN	Quaker	HOD
36	Anna Gurney	NaN	NaN	Quaker	HOD
37	Anna Gurney	NaN	NaN	Quaker	HOD
38	Samuel Gurney	NaN	NaN	Quaker	HOD
39	Samuel Gurney	NaN	NaN	Quaker	HOD
40	John Barton Hack	NaN	NaN	Quaker	HOD
41	John Hodgkin	NaN	NaN	Quaker	HOD
42	John Hodgkin	NaN	NaN	Quaker	HOD
43	Luke Howard	NaN	NaN	Quaker	HOD
44	Luke Howard	NaN	NaN	Quaker	HOD
45	William Howitt	NaN	NaN	Quaker	HOD
46	Andrew Johnston	NaN	NaN	Quaker	HOD
47	Joseph Pease	NaN	NaN	Quaker	HOD
48	Joseph Whitwell Pease	NaN	NaN	Quaker	HOD

6.38 Quaker CEDA members not in Case Study Three

quaker_not_hod

	Name	religion_name	ceda_name	first_year	last_year
0	William Spicer Wood	Quaker	APS	1864	1867
1	William Spicer Wood	Quaker	ASL	1863	1871
2	William Spicer Wood	Quaker	AI	1863	1871
3	William Wilson	Quaker	APS	1838	1865
4	William Wilson	Quaker	ASL	1865	1866
...
634	Joshua Wilson	Quaker	APS	1860	1860
635	F Woodhead	Quaker	APS	1861	1862
636	W Woolston	Quaker	APS	1861	1861
637	Francis Wright	Quaker	APS	1838	1838
638	S W Wright	Quaker	APS	1861	1861

639 rows x 5 columns

P7 Chapter 7 Project Seven presentation

File name: jnb_hddt_qsra_september_2021

7.1 An Evidence Based Prosopographical study of 3000 activists 1830-1870 and the 600 Quakers amongst them

Subject:

My own research has revealed the extensive social connectivity between the roughly 600 members of a 'Quaker Led Network (QLN)' and their involvement within a community of roughly 3000 persons, spread across four organisations in Britain active between 1830 and 1870, which the QLN network helped to set up and staff. I call these, the 'Centres for the Emergence of Discipline of Anthropology in Britain' (CEDA).

Question 1:

Can the model reveal the networks between the members in the five organisations that comprise the CEDA? This question is important because it resolves a wider and current uncertainty over the origins of the discipline of anthropology in Britain and the extent of Quaker involvement.

Methodology:

I have designed, built and I am now using a suite of open-source and reproducible relational database technologies and digital analytic tools to visualise and scrutinise the entire community of some 3000 activists over 40 years (1830-1870), picking out the Quakers amongst them so that the community can be explored at both group and individual level. I am able to model the 'connected' relationships between the individual members of the CEDA through time, including kinship, education, occupations, locations and organisations.

Question 2:

Can the model examine Quaker-to-Quaker relationships and how these relationships supported the Quaker members of the CEDA during the forty years of its life?

Question 3:

Can the model reveal the key networking role played by the Quaker Thomas Hodgkin MD (1798–1866) from the beginnings of the CEDA in 1830 up to his death in 1866?

7.2 This is a code cell


```
# First we call up the python packages we need to perform the analysis:

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
plt.rc('figure', figsize=(20, 10))
from IPython.display import set_matplotlib_formats
set_matplotlib_formats('png', 'pdf')
from operator import itemgetter
import networkx as nx
from networkx.algorithms import community #This part of networkx, for commun
import nbconvert
import csv

# 
```

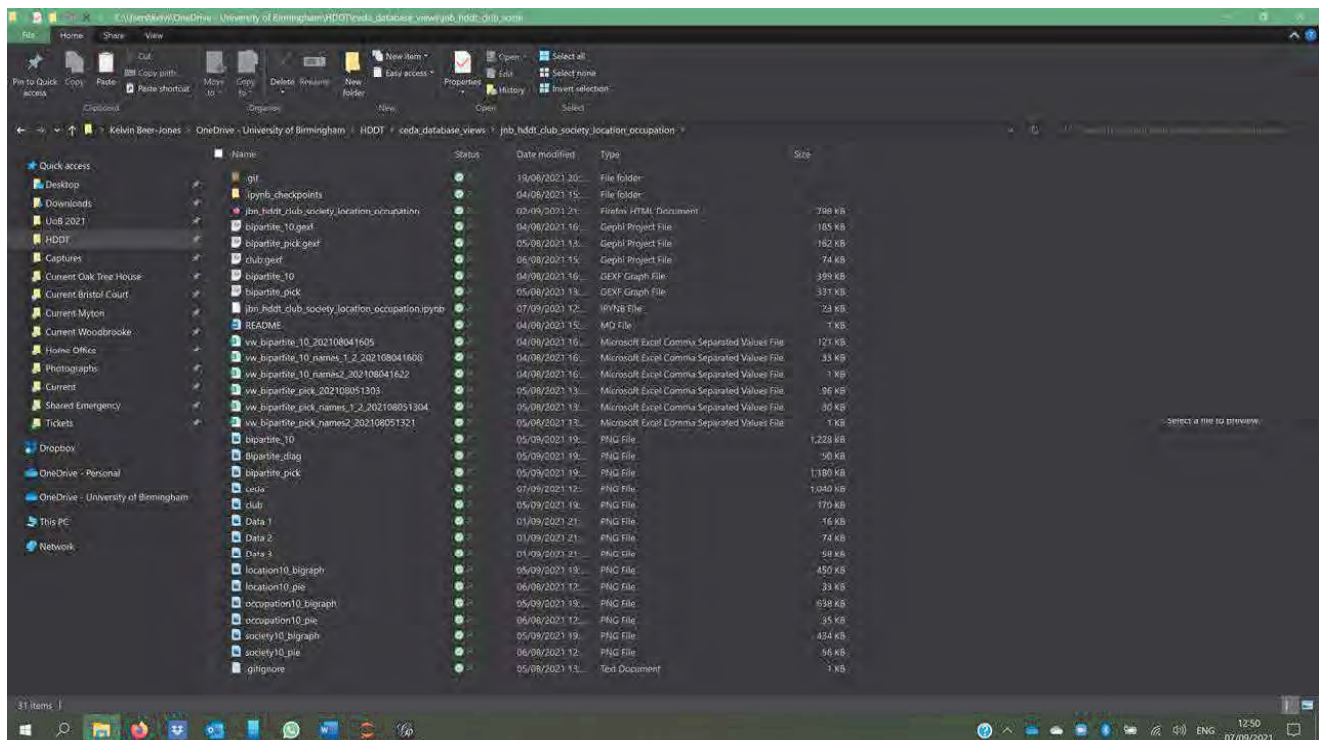
```
-----
ModuleNotFoundError                                Traceback (most recent call last)
Cell In[1], line 5
      3 import pandas as pd
      4 import numpy as np
----> 5 import matplotlib.pyplot as plt
      6 plt.rc('figure', figsize=(20, 10))
      7 from IPython.display import set_matplotlib_formats

ModuleNotFoundError: No module named 'matplotlib'
```

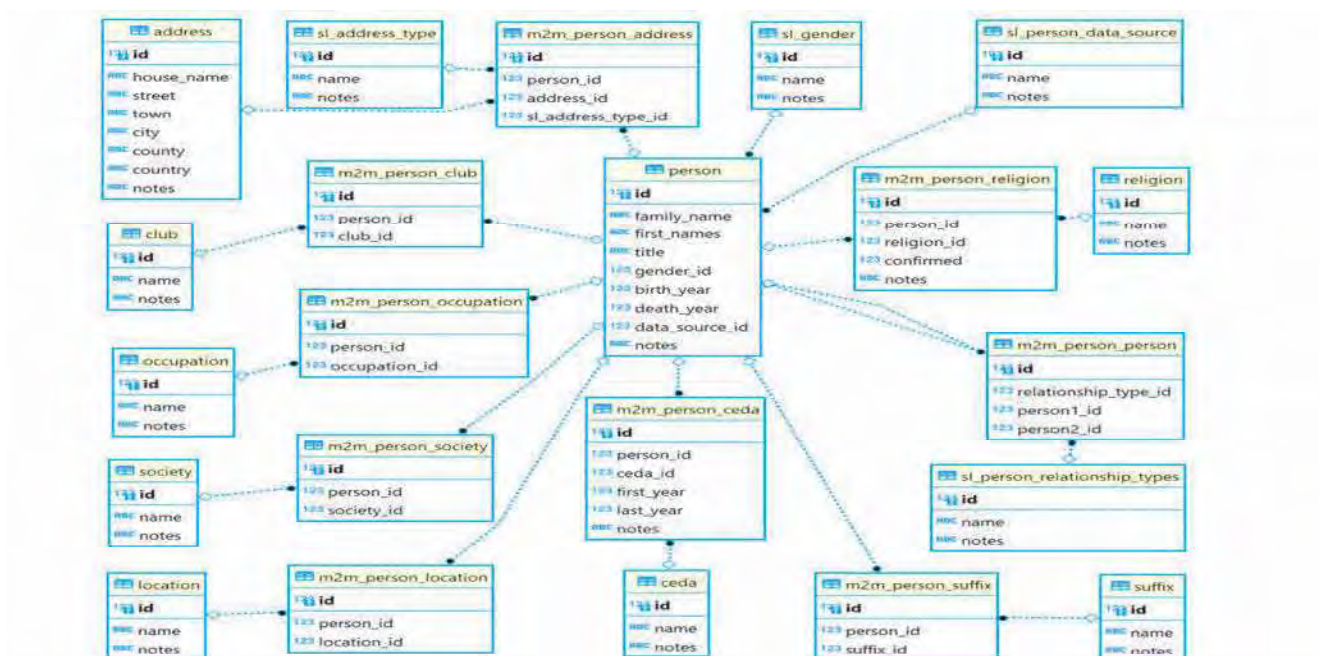
7.3 These are the resources in my container for this exercise

Resource containers are also GitHub repo's facilitating granular version control of all changes

made to any resources

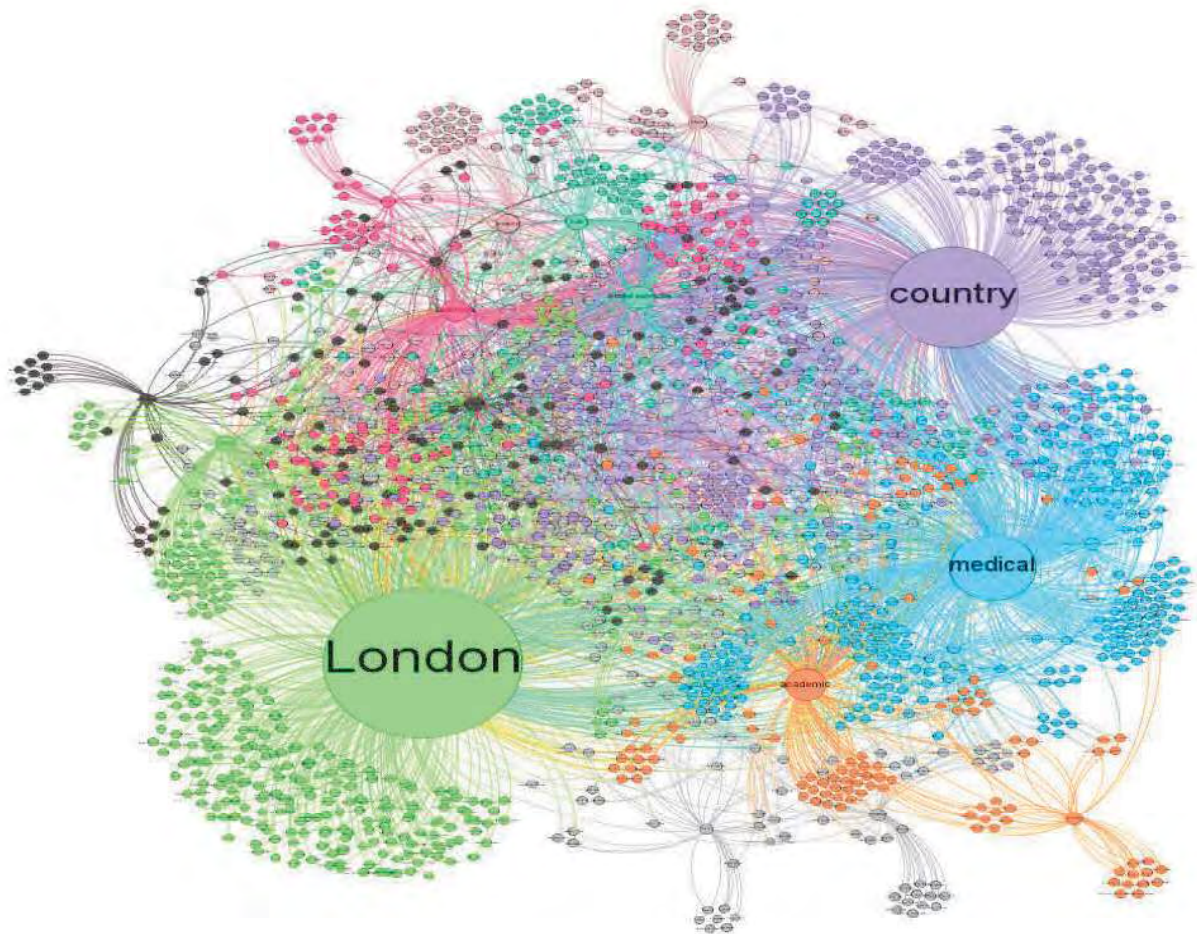


7.4 This is the structure of the SQLite database (ERD)



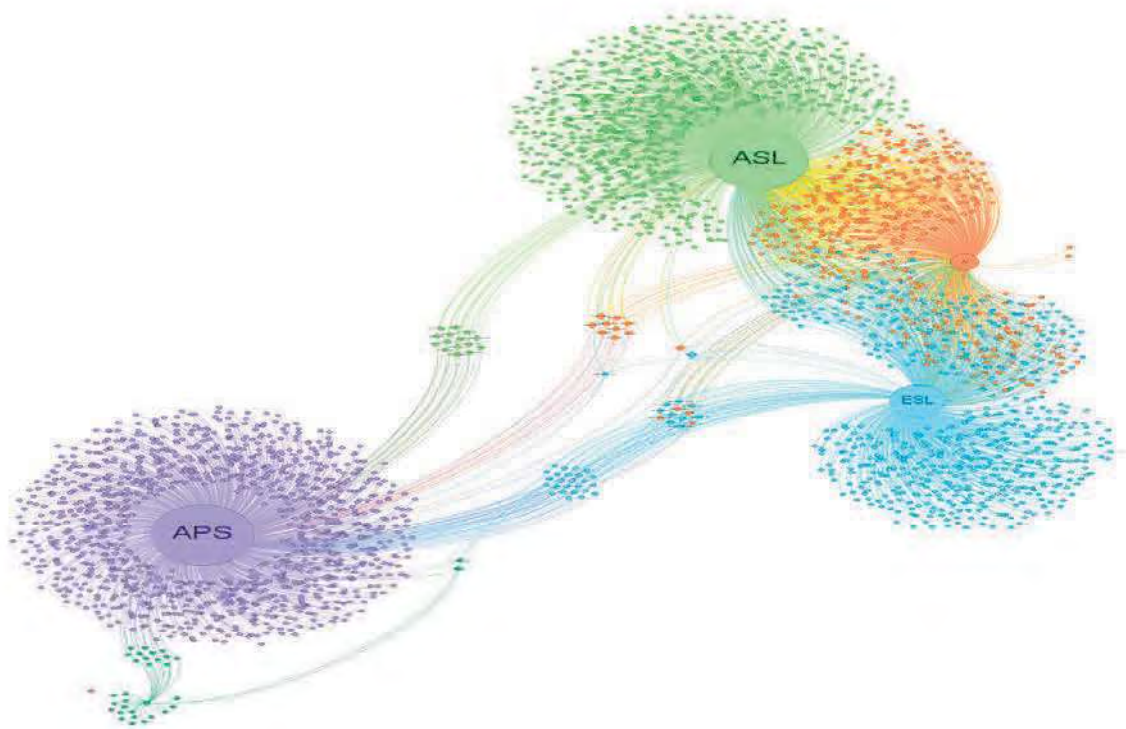
7.5 Relationships (other than CEDA)

present in the data



This graph shows all of the popular bigraph data in the database. Including all data would result in a 'hairball' where data would be too dense to be capable of analysis at this (the highest) level. (See Most popular entities section below)

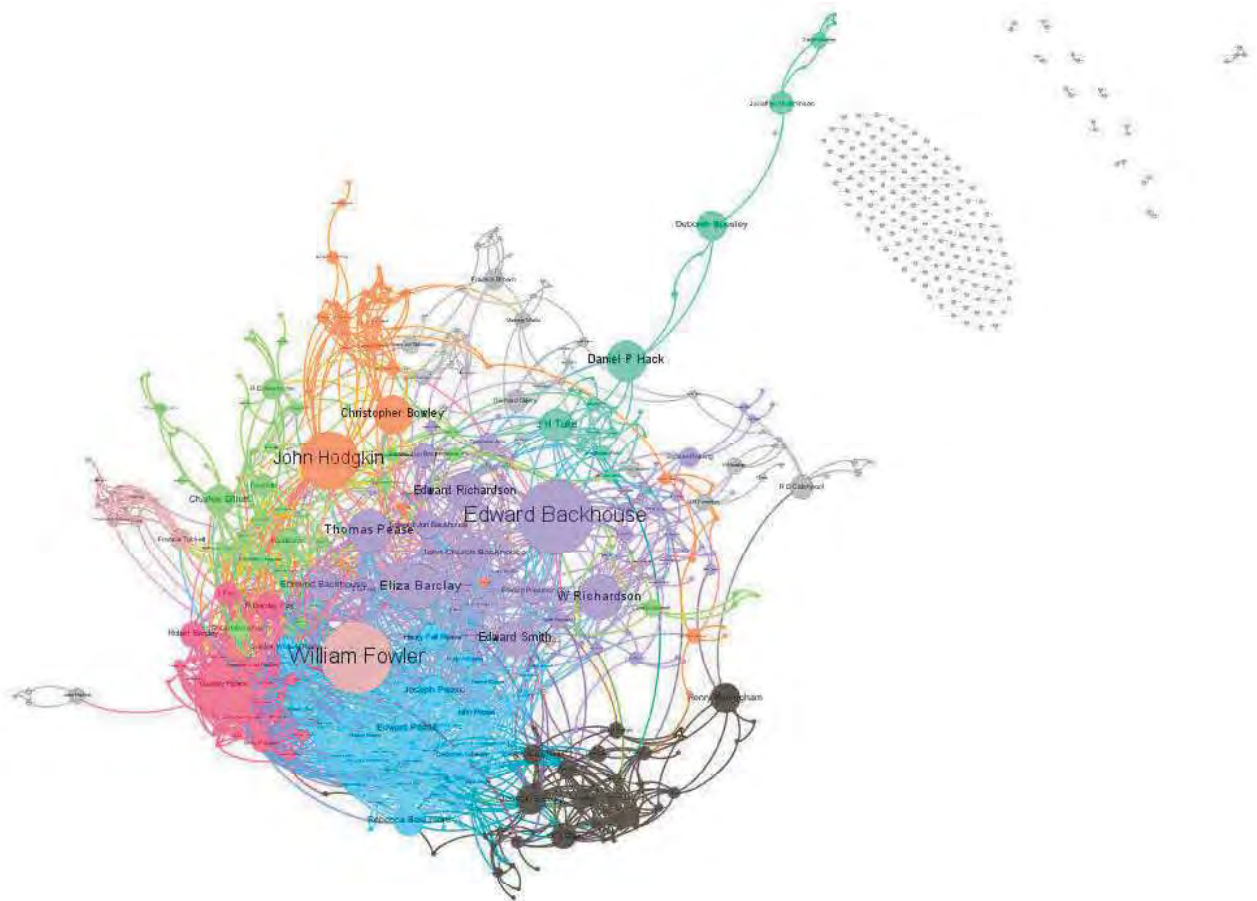
7.6 All persons are members of at least one



Society	abv.	Dates
Quaker Committee on the Aborigines*	QCA	1832/37 - 1846
Aborigines Protection Society	APS	1837 - 1919
Ethnological Society of London	ESL	1843 - 1871
Anthropological Society of London	ASL	1863 - 1871
Anthropological Institute	AI	1843 - 1871
London Anthropological Society**	LAS	1873 - 1874

- Origin Society included in this project but not recognised by RAI. ** not included in this project (beyond 1871 cut off date).

7.7 Quakers and their relationships



7.8 Working with a variety of datatables

A 'complete' dataset

Would be one like this, where all of the data can be contained within a perfect rectangular block of cells ('containers') and every container contains only one data item and every data item can be located by the coordinates 'Row n, Column n'

Table	Col1	Col2	Col3	Col4	Col5
Row1	A	B	C	D	E
Row2	F	G	H	I	J
Row3	K	L	M	N	O
Row4	P	Q	R	S	T
Row5	U	V	W	X	Y

An 'incomplete' dataset

When historical data is used often some data is missing (permanently lost) and the HDDT is able to accept 'Incomplete' datasets. The HDDT does not lose functionality because of the incomplete nature of much historical data.

Table	Col1	Col2	Col3	Col4	Col5
Row1		B	C	D	E
Row2	F		H	I	J
Row3	K	L		N	O
Row4	P	Q	R		T
Row5	U	V	W	X	

An 'irregular' dataset

The HDDT has been designed to accept Irregular datasets. The surviving evidence of the past is not only often Incomplete, but also frequently Irregular, where multiple datasets have different dimensions. (Either because the data in itself is intrinsically different or because different data collectors use different cataloguing methods).

1	A		1.0		
2	B		2.1	Cat	Q
3	C		3.2	Dog	W
4	D	fff	4.3	Fish	E
5	E	ggg			R
6	F	hhh			T
7	G	iii			Y

For the HDDT a qualifying dataset is a data set of any dimensions, complete, incomplete or irregular. The only requirement is that all datasets must contain a **single** common containing one universally shared data item. The HDDT requires all data sets to contain datatables that can be referenced to a PERSON (Name) in one of its rows.

Conflicts between dataset Person (Name)'s are resolved by adopting in this project by nominating the 'RAI dataset' as the 'Authority Index'. With careful matching of Person

(Name)'s found in other datasets, the RAI naming rule applies throughout.

7.9 Introduction to bigraph analysis

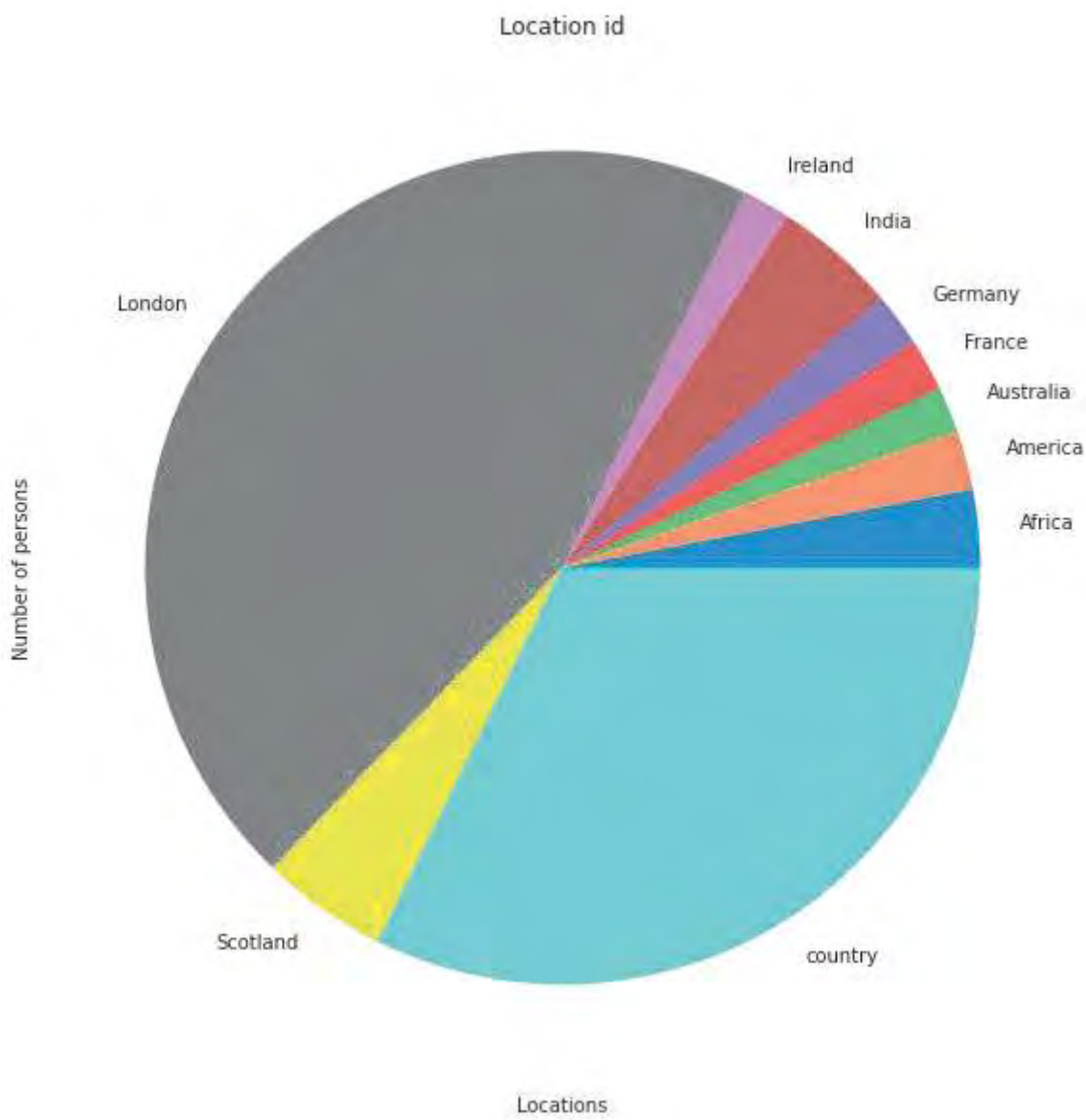
Bigraph data statistics

Table	Rows	Columns	Exc?
ceda	6	1	Yes
person_ceda	3894	4	Yes
club	68	1	*
person_club	323	2	*
location	83	1	
person_location	2061	2	
occupation	93	1	
person_occupation	1883	2	
society	260	1	
person_society	1238	2	

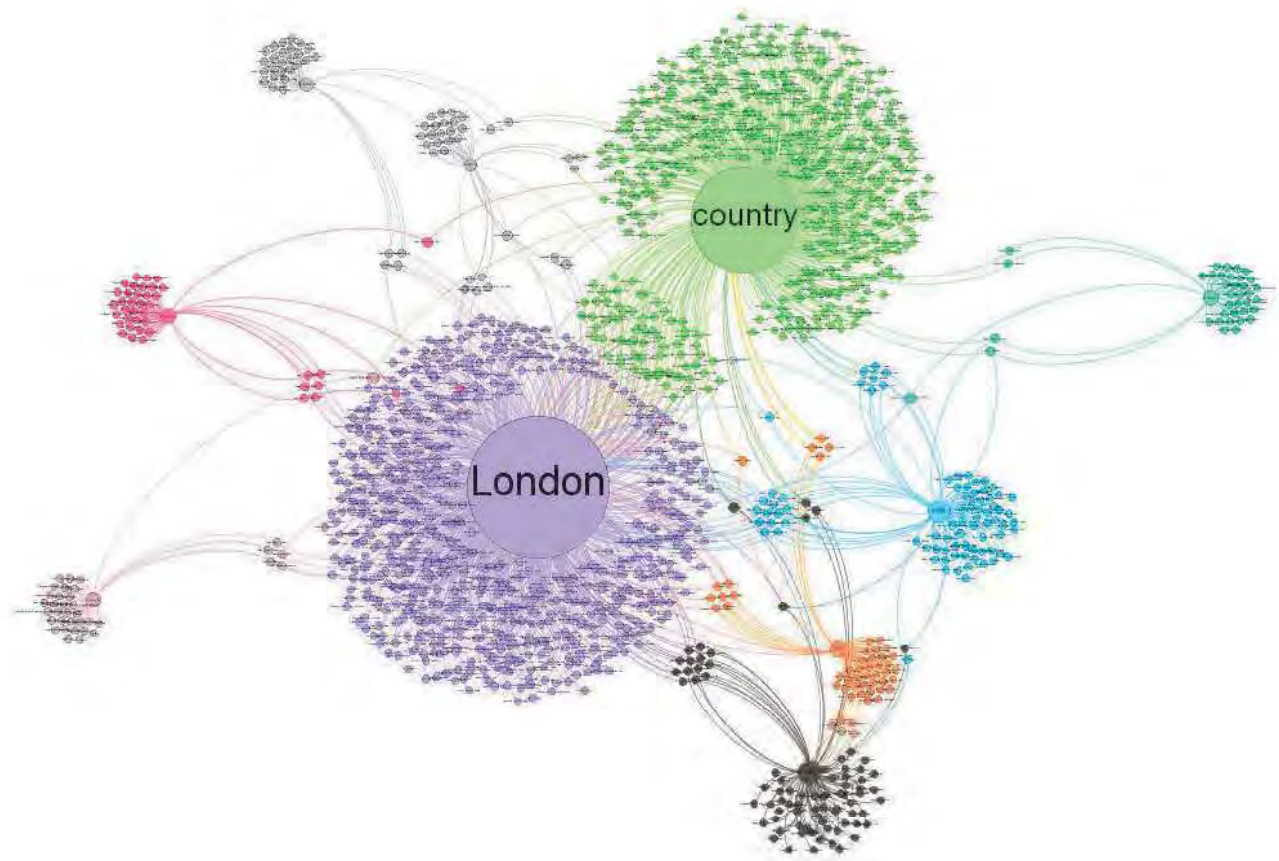
- Due to low levels of poulation of 67 other clubs only the Athenaeum Club is included in analysis

7.10 Locations

Top 10 locations pie chart



Top 10 locations bigraph



We can see that London and 'country' (sic) are the most populated locations. Because the 'country' location is an aggregate (and not a specific location) we can think of London and 'country' as a twin centre. Within the twin centre we can see the members of both London and 'country' locations and that the members of each are highly networked. We can also see that the London location contains many members who have no association with any other group (including 'country'). London 1830 - 1870, was densely populated and so it is possible that members of the London location had other modes of association. Because the 'country' location is an aggregate we cannot make the same analysis to the same extent, it is possible that many members in (say) Newcastle had no association with other members in (say) Bristol. We can see the large group of members who were members of both London and 'country' locations. It is highly likely that these members served as conduits of communication and group cohesion. It is interesting to note that only 3 members of this London and 'country' group were members of groups outside of the twin centre.

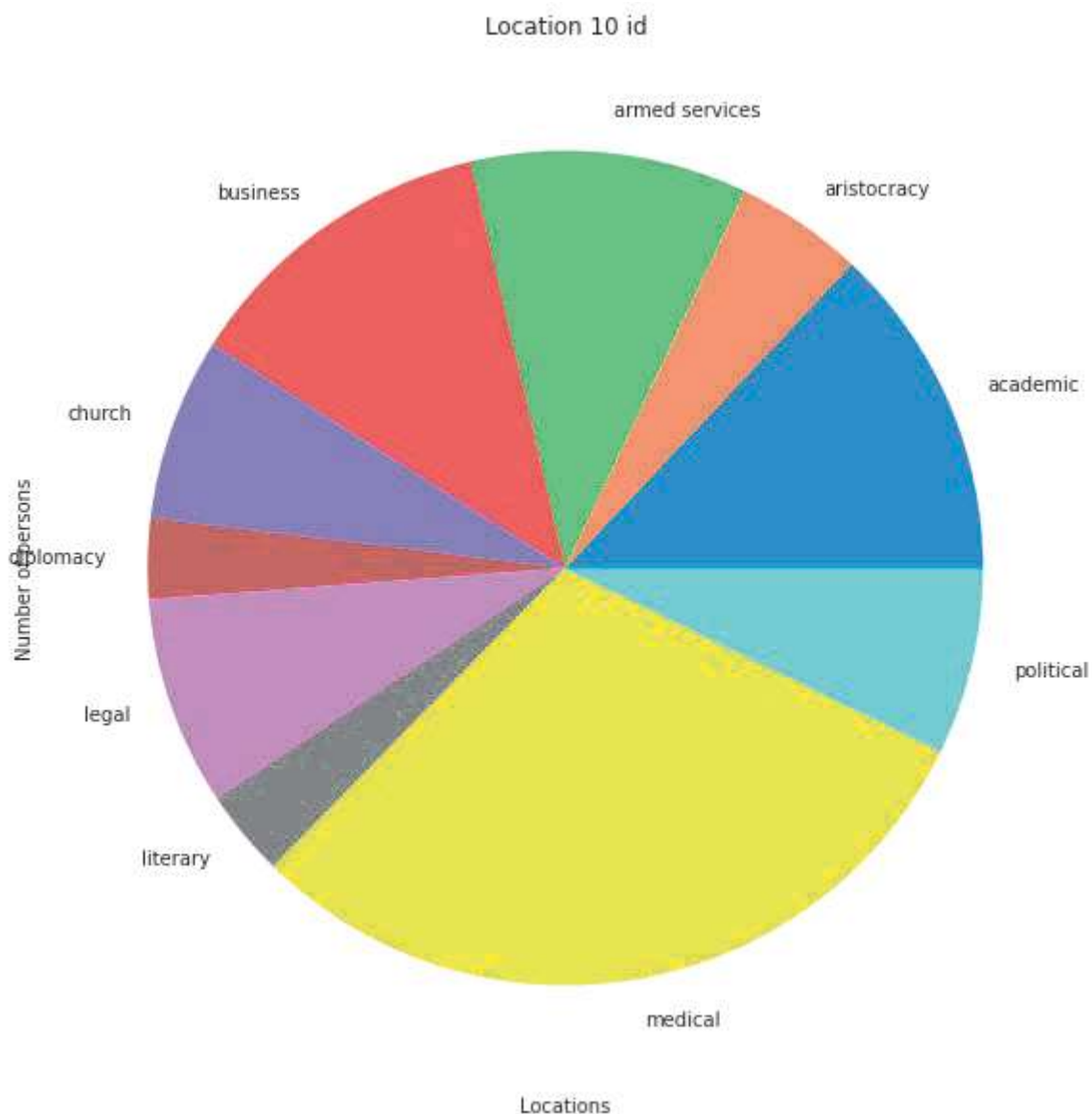
Eight other location each have a membership of around 30 members (we can call these the satellites), all of the satellite groups relate directly to the twin centre with very few members associated with more than one satellite location.

Australia and Ireland have associations with both London and 'country'. The German location is most closely associated with the 'country' group. All of the other locations are strongly associated with the London location.

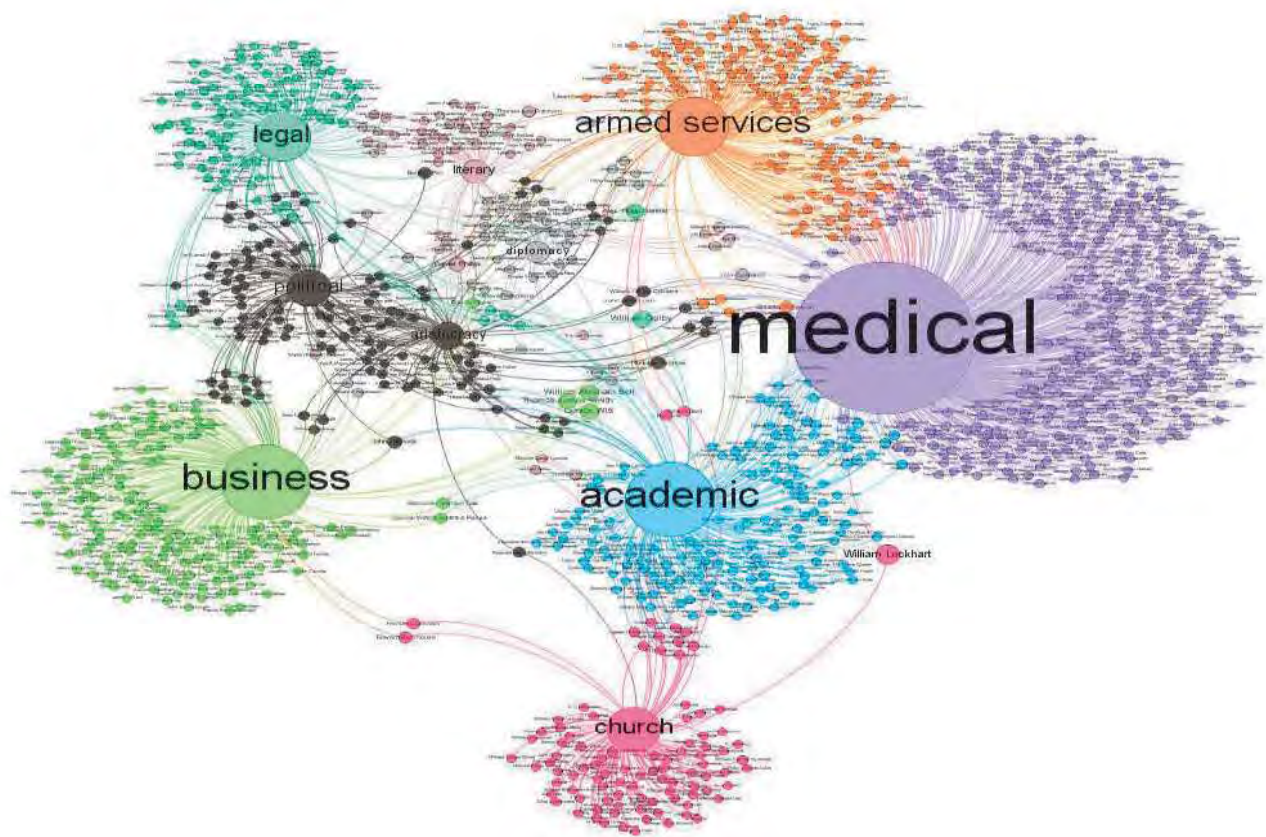
Germany (far right) is the location least associated with London. Alex Nidda Genthe is the only member from Germany who is also a member of the London location. Friedrich Max Muller, Frederick Augustus Haverick and Gustav Oppert each network with 'country' members. William Wilson Hunter is the only 'country' member who also appears in the Germany location. He and Gustav Oppert also have a location connection with India.

7.11 Occupations

Top 10 occupations pie chart



Top 10 occupations bigraph



We can see that 'medical', 'academic' and 'armed services' together account for half of the members by occupation. We can also see that the largest three occupational categories each contain many members who have no association with any other occupational group. We can see that the medical categories contain many members who are also members of the other two principal categories ('academic' and 'armed services'). It is highly likely that these members served as conduits of communication and group cohesion amongst the three principal occupational categories.

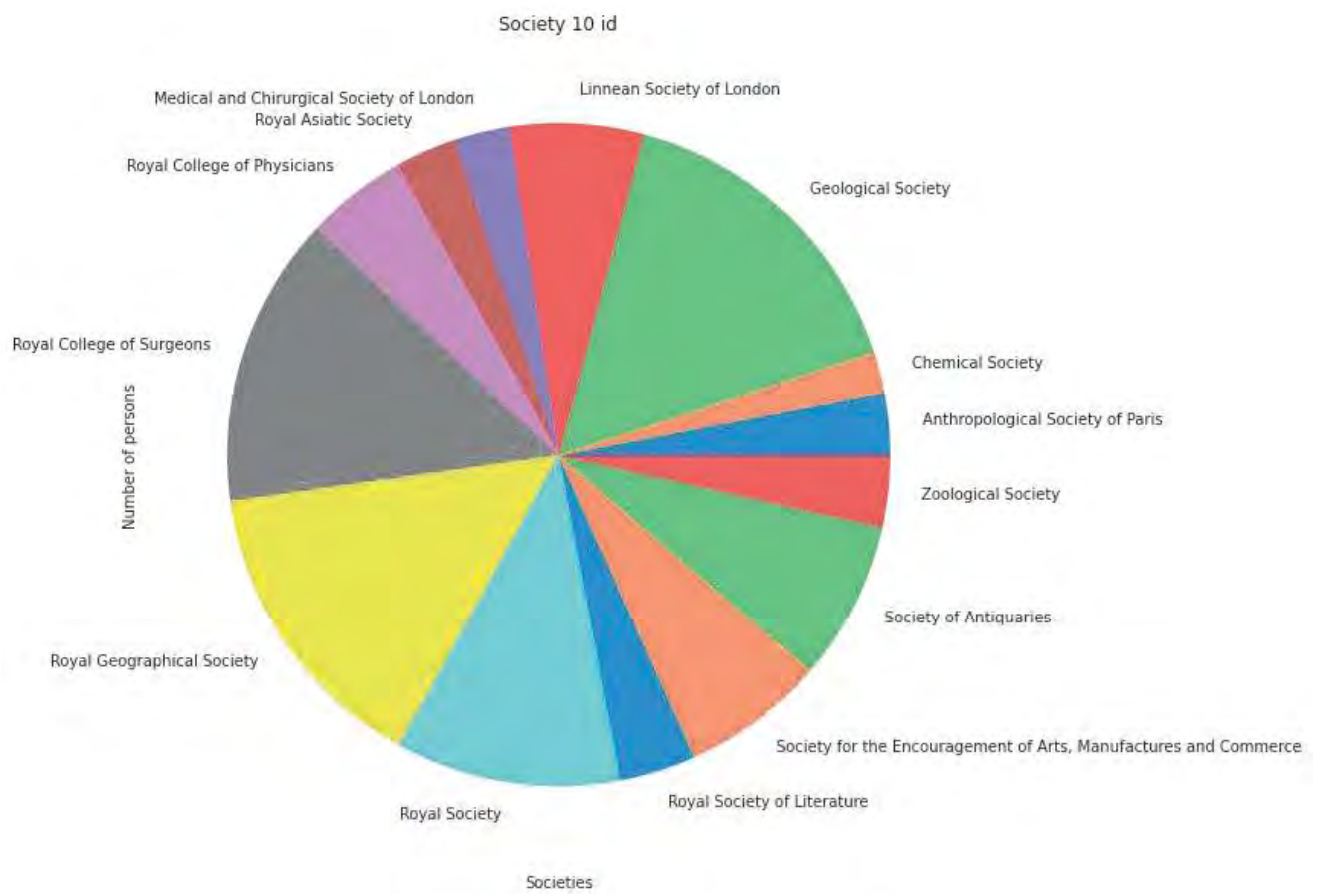
Seven other occupations each have a range of members with literary the lowest and business the highest. All of the satellite groups relate directly to the triple centre with many members also associated with more than one other satellite occupation.

It is surprising the least networked occupation is 'church' and perhaps less so that 'business' and 'legal' are highly networked.

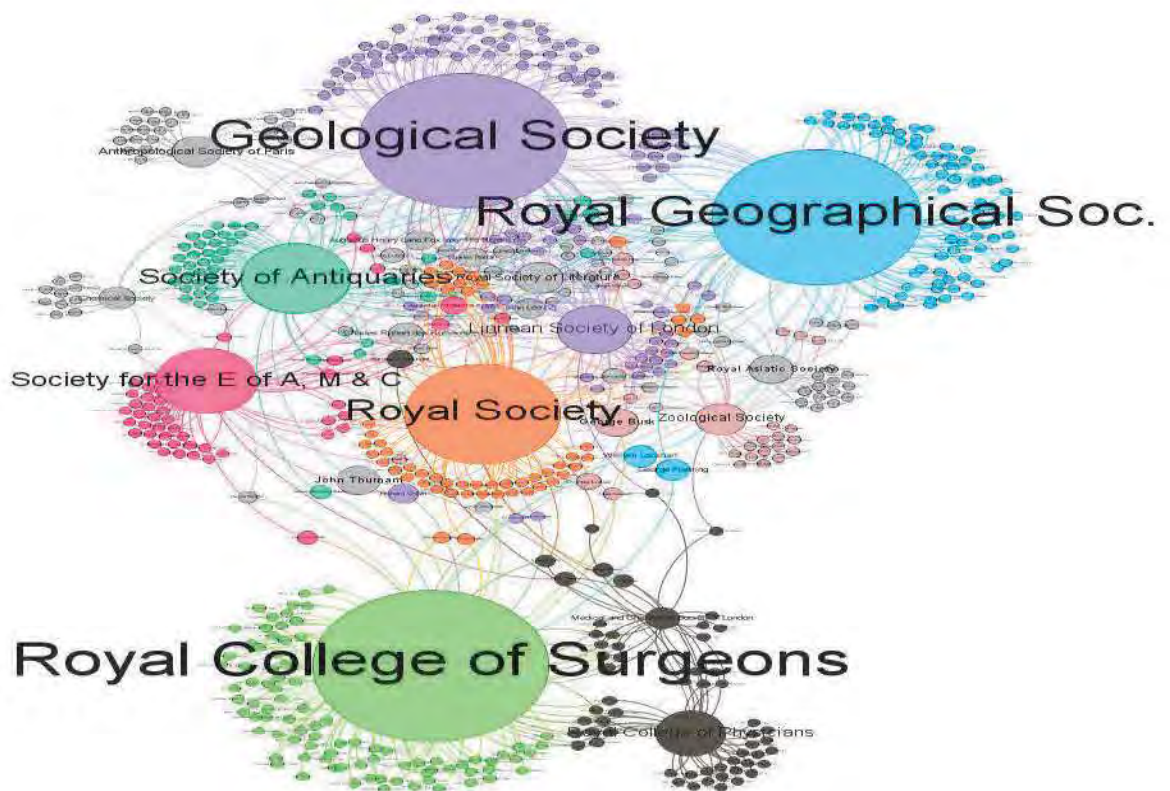
Several individuals form a web of interconnectedness between the members occupations.

7.12 Societies

Top 10 societies pie chart



Top 10 societies bigraph



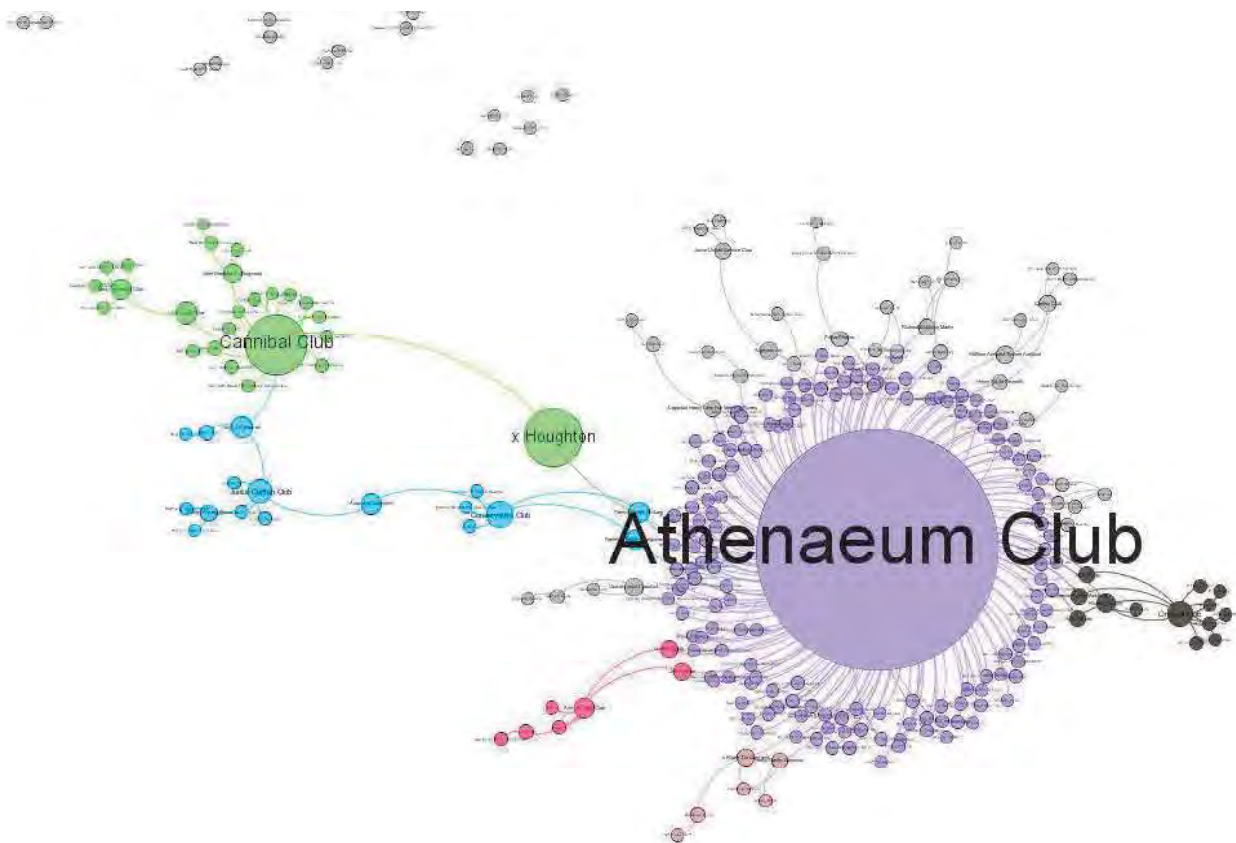
We can see that 'Geological Society' and the 'Royal Geographical Society' together account for a significant number of members by society. The 'Royal College of Surgeons', the 'Medical and Chirurgical Society' and the 'College of Physicians' form the next largest cluster of memberships of societies. These two clusters each contain many members who have no association with any other society. We can see that the medical group and the geographical group have few members in common. The 'Royal Society' and the 'Linnean Society' in the centre have between them the greatest level of networking amongst all of the societies. It is highly likely that these members served as conduits of communication and group cohesion amongst the two principal society groups.

Many other societies have a range of members all of whom are highly interconnected. All of the satellite groups relate most closely to the 'Royal Society' and the 'Linnean Society' rather than to the two larger clusters. Many members of the smaller satellite societies are also associated with more than one other satellite occupation.

It is surprising the least networked occupation is the 'Royal College of Surgeons' and perhaps less so that the 'Geological Society' and the 'Royal Geographical Society' are highly networked.

Several individuals form a web of interconnectedness between the members of societies.

7.13 Cubs



Clubs will not be analysed in this project but the Athenaeum club can be used as an attribute (because it is a singularity).

7.14 Most popular bipartite networks combined

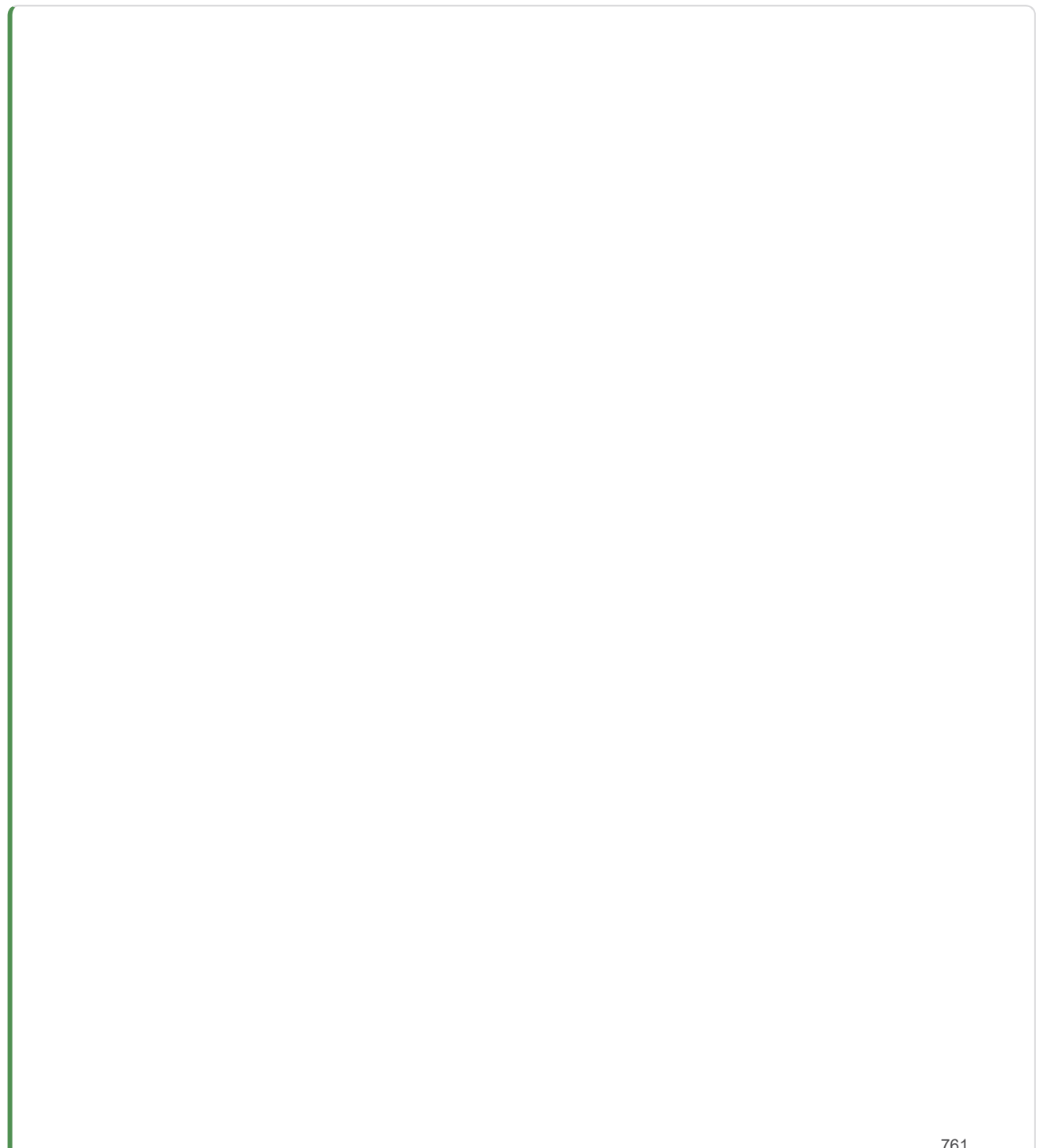
1850 members of the community are recorded as members of 35 popular entities (Locations, occupations, societies and the Athenaeum Club). These entities make a sphere of popular interest graph where meetings between members concerning the CEDA may have taken place, equally they may also be places where members might meet up only infrequently or informally. The visual analysis of connectivity between members in single societies and between members of multiple societies indicates the extent that the community is societally connected. The 1850 make up 60% of the entire community.

The graph at the head of this notebook shows the 1850 distributed by popular entity membership with the connectivity between them reflected in those members who are associated with more than one entity.

7.15 Iterative Section 1 - (This is an iterative workbook)

As can be seen in the illustrative graph above which has been produced in Gephi by the code cells below to provide an initial overview of the data and its distribution, the graph might be made more meaningful if it did not include societies sparsely populated.

The code cell below and the code cells in the section Iterative Section 2 (below) have therefore been designed so that a second run through the workbook can be made where the second run uses data that excludes low populated occupations identified in the first run through.



```

# Second we call up the csv files generated from the SQL database that contain
# locations and the community members associated with locations. As well as
# we produce a 'node_names' file and a tuples file of edges_attributes to generate
# produce GefX files for Gephi.

# We can run the code cell twice, first with all data and once all data has been
# and a decision made to exclude 'noise' the code block can be run again with
# csv files that exclude low populated locations.

bipartite_10 = pd.read_csv ('vw_bipartite_10_names2_202108041622.csv')
bipartite_pick = pd.read_csv('vw_bipartite_pick_names2_202108051321.csv')

# Use these csv files in the 'with open' statements below to generate bipartite
#names = pd.read_csv ('vw_bipartite_10_names_1_2_202108041606.csv')# For node names
#tuples = pd.read_csv ('vw_bipartite_10_202108041605.csv')# For edges.csv

# Use these csv files in the 'with open' statements below to generate location
bipartite_pick_names = pd.read_csv ('vw_bipartite_pick_names_1_2_202108051304.csv')
bipartite_pick_tuples = pd.read_csv ('vw_bipartite_pick_202108051303.csv')

with open('vw_bipartite_pick_names_1_2_202108051304.csv', 'r') as nodecsv:
    nodereader = csv.reader(nodecsv) # Read the csv
    nodes = [n for n in nodereader][1:]# Retrieve the data (using Python list slicing)
    # to remove the header row
    node_names = [n[0] for n in nodes] # Get a list of only the node names

with open('vw_bipartite_pick_202108051303.csv', 'r') as edgecsv: # Open the
    edgereader = csv.reader(edgecsv) # Read the csv
    edge_list = list(edgereader) # Convert to list, so can iterate below in

# Create empty arrays to store edge data and edge attribute data
edges = []
edges_attributes = []

# Fill the arrays with data from CSV
for e in edge_list[1:]:
    edges.append(tuple(e[0:2])) # Get the first 2 columns (source, target)
    # not used this time. edges_attributes.append(tuple(e[2:4]))
    # Get the 3rd and 4th columns (first_year, last_year) and add to array

edge_names = [e[0] for e in edges] # Get a list of only the edge names

```

7.16 Listing out the data

```

# List out the societies to be analysed

# bipartite_10

```



```
# List out the community members who have been associated with at least one  
# names
```

```
# Finally list out the tuples of members and societies  
# (Note – some members are associated with more than one society)  
  
# tuples
```

7.17 Iterative Section 2 - prepare the data for rendering as a graph in Gephi

Caution - this section depends on the selections made under 'Iterative Section 1' above

If the initial analysis suggests that a more insightful visualisation might be made by refining the data to be analysed, return to the database and make a new Nodes (Names) csv file and a new Tuples csv file containing only well populated groups. Then return to Iterative Section 1 code block in the workbook and replace the csv files in the 'with open' code lines with the refined datasets. Finally reset the `nx.write_gexf (xxx.gexf) xxx` statement to a new file name. Then run all code blocks again and make a more insightful gexf file. Use that to produce an improved network graph for Stage 2 analysis.

Warning. - Ensure that the statement '`nx.write_gexf`' in the last code cell in this section points to a new output file for Gephi. (e.g., `G, 'xxxx_10.gexf'`) Failure to set this value correctly will result in the previously generated .gexf file being overwritten instead.

```
print("Nodes length: ", len(node_names))  
print("Edges length: ", len(edges))  
# not used this time. print("Edges attributes length: ", len(edges_attributes))
```

```
Nodes length: 1683  
Edges length: 3261
```

```
# First check that the data is correctly formatted

print("First 5 nodes:", node_names[0:5])
print("First 5 edges:", edges[0:5])
# not used this time. print("First 5 edges attributes:", edges_attributes[0

# The output will appear below this code cell.
```

```
First 5 nodes: ['A Mackintosh Shaw', 'A , jun Ramsay', 'A A Stewart', 'A B
First 5 edges: [('Arthur William A Beckett', 'London'), ('Andrew Mercer Adam
```

```
# We use NetworkX to build the graph data into a table

G = nx.Graph()
G.add_nodes_from(node_names)
G.add_edges_from(edges)
print(nx.info(G))
```

```
Name:
Type: Graph
Number of nodes: 1683
Number of edges: 3261
Average degree: 3.8752
```

```
# Finally we can write a gexf file which will be placed in the root director
# We can then open the file in Gephi and visualise the network.

#nx.write_gexf(G, 'bipartite_pick.gexf')
```

7.18 Stage 2 - Bipartite analysis with 'noise' removed

We now re-run the code to generate a new gexf file for gephi. We use the refined pair of nodes (Names) and Tuples files generated in the SQL database that include only the top 17 groups.

```
bipartite_pick
```

	Name
0	Athenaeum Club
1	Chemical Society
2	Geological Society
3	Linnean Society of London
4	London
5	Medical and Chirurgical Society of London
6	Royal College of Physicians
7	Royal College of Surgeons
8	Royal Geographical Society
9	Royal Society
10	academic
11	armed services
12	business
13	country
14	diplomacy
15	medical
16	political

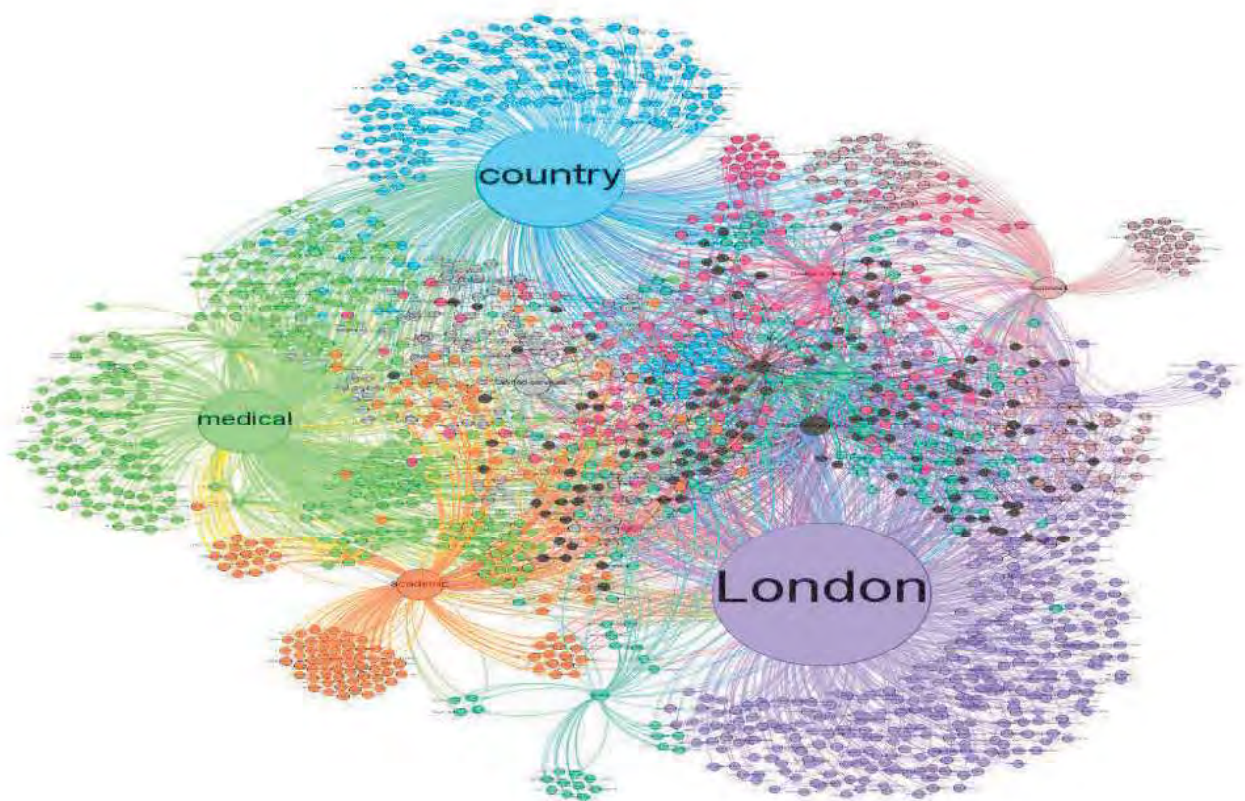
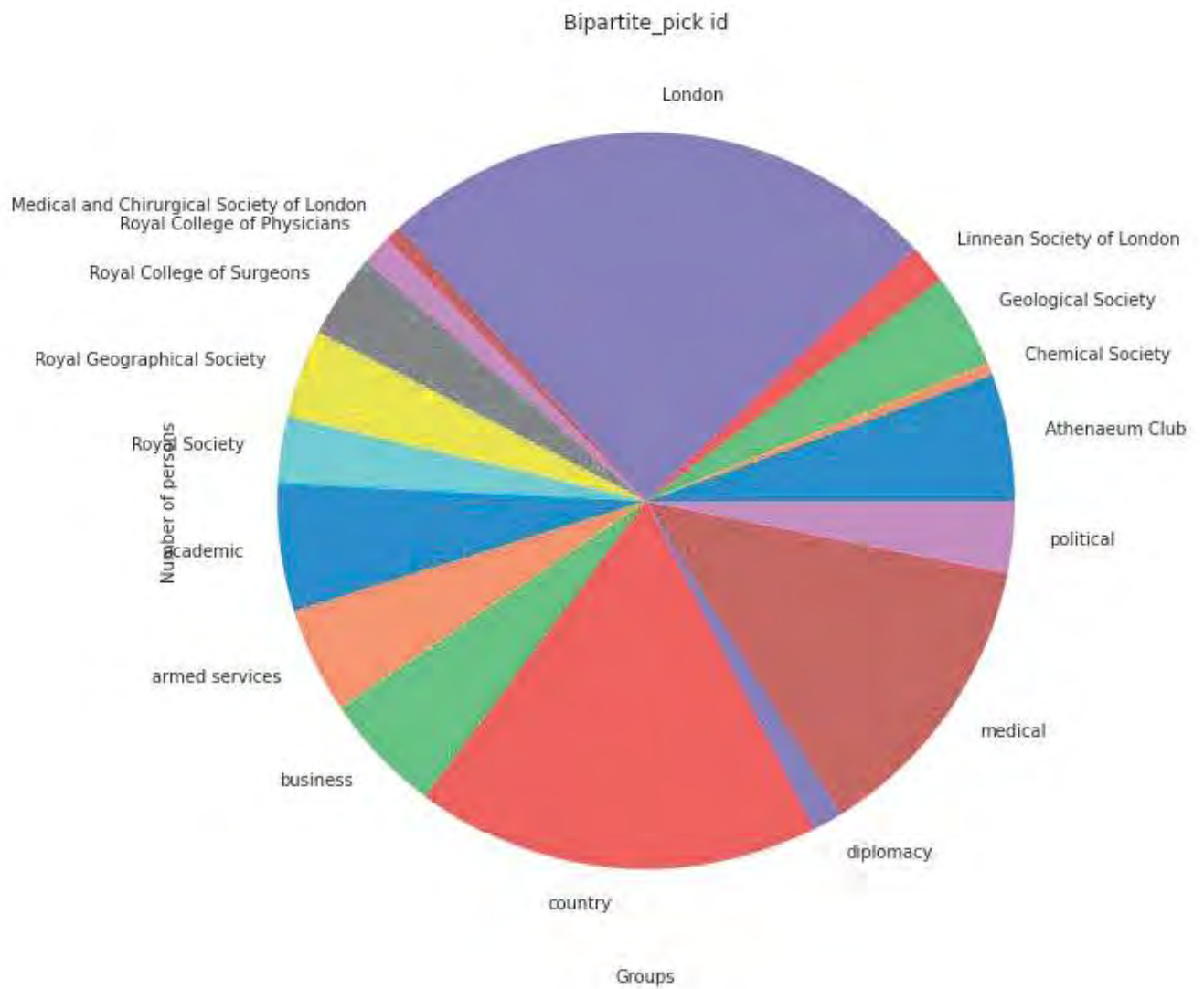
bipartite_pick_tuples

	Source	Target
0	Arthur William A Beckett	London
1	Andrew Mercer Adam	armed services
2	Andrew Mercer Adam	country
3	Andrew Mercer Adam	medical
4	William Adam	political
...
3256	James A Youl	London
3257	James A Youl	business
3258	Robert Younge	Linnean Society of London
3259	Robert Younge	country
3260	Arthur de Zeltner	diplomacy

3261 rows x 2 columns

7.19 Simplified graphs can be analysed more easily

```
bipartite_pick_tuples.groupby('Target')['Source'].nunique().plot(kind='pie')
plt.title ("Bipartite_pick id")
plt.xlabel ("Groups")
plt.ylabel ("Number of persons")
plt.show()
```



The community members most well connected (60%) are densely networked indicating that the CEDA members are able to bring to the task of developing the discipline of

anthropology in Britain considerable shared skills, information and knowledge.

List of References

- Aborigines Protection, Society. 1837. *Report of the Parliamentary Select Committee on Aboriginal Tribes (British settlements) reprinted with comments*. London: William Ball.
- Aborigines Protection Society. 1838. *The First Annual Report of the Aborigines Protection Society*. (London).
- . 1839. *The Second Annual Report of the Aborigines Protection Society*.
- An Address of Christian counsel and caution to emigrants to newly-settled colonies*. 1841. Edited by Committee London Yearly Meeting . Aborigines. Vol. Accessed from <https://nla.gov.au/nla.cat-vn1027098Rex> Nan Kivell Collection ; NK7097. London: Harvey and Darton.
- Allemang, Dean, and Jim Hendler. 2011a. *Semantic web for the working ontologist: effective modeling in RDFS and OWL*. 2nd ed.: Boston: Morgan Kaufmann.
- . 2011b. "What is the Semantic Web?" In *Semantic Web for the Working Ontologist* edited by Dean Allemang and Jim Hendler. Boston: Morgan Kaufmann.
- Anderson, Sheila, Tobias Blanke, and Stuart Dunn. 2010. "Methodological commons: arts and humanities e-Science fundamentals." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368 (1925): 3779-3796. <https://doi.org/doi:10.1098/rsta.2010.0156>.
- Atici, Levent, Sarah Witcher Kansa, Justin Lev-Tov, and Eric C Kansa. 2013. "Other People's Data: A demonstration of the imperative of publishing primary data." *Journal of Archaeological Method and Theory* 20 (4): 663-681. <https://doi.org/https://doi.org/10.1007/s10816-012-9132-9>.
- Baker, Thomas, Emmanuelle Bermès, Karen Coyle, Gordon Dunsire, Antoine Isaac, Peter Murray, Michael Panzer, Jodi Schneider, Ross Singer, and Ed Summers. 2011. "Library Linked Data Incubator Group Final Report: W3C Incubator Group Report 25 October 2011." <https://www.w3.org/2005/Incubator/ld/XGR-ld-20111025> (Accessed 21 March 2025).
- Bawden, David. 2015. "Introduction to Information Science Chapter One." Facet Publishing.

- Bekiari, Chrysoula, George Bruseker, Martin Doerr, Christian-Emi Ore, Stephen Stead and Athanasios Velios. 2021. Definition of the CIDOC Conceptual Reference Model. CIDOC CRM Special Interest Group.
- Bell, Mark, and Sonia Ranade. 2015. "Traces Through Time: a case-study of applying statistical methods to refine algorithms for linking biographical data." BD.
- Benardou, Agiatis. 2010. "Understanding the Information Requirements of Arts and Humanities Scholarship." *International Journal of Digital Curation* 5 (1). 5 (1). <https://doi.org/https://doi.org/10.2218/ijdc.v5i1.141>.
- Benardou, Agiatis, Panos Constantopoulos, Costis Dallas, and Dimitris Gavriliis. 2010. "A conceptual model for scholarly research activity." *Digital Curation Unit-IMIS, Athena Research Centre*. <https://www.ideals.illinois.edu/items/14918> (Accessed 21 March 2025).
- Berg, Kristi L, Tom Seymour, and Richa Goel. 2013. "History of Databases." *International Journal of Management & Information Systems (IJMIS)* 17 (1): 29-36.
- Berners-Lee Tim, Hendler James, and Lassila Ora. 2001. "The Semantic Web." *Scientific American*. http://www.sciam.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21.
- Bianchini, Carlo. 2022. "The entities of the IFLA-LRM, RiC-CM and CIDOC-CRM models in the semantic web." *JLIS.it* 13 (3): 63-75. <https://doi.org/10.36253/jlis.it-482>.
- Bianchini, Carlo, Stefano Bargioni, and Camillo Carlo Pellizzari di San Girolamo. 2021. "Beyond VIAF: Wikidata as a Complementary tool for authority control in libraries." *Information Technology and Libraries* 40 (2). <https://doi.org/https://doi.org/10.6017/ital.v40i2.12959>.
- Blaney, Jonathan, Jane Winters, Sarah Milligan, and Martin Steer. 2021. *Doing digital history: a beginner's guide to working with text as data*. Manchester: Manchester University Press.
- Blaxill, Luke. 2023. "Why do Historians Ignore Digital Analysis? Bring on the Luddites." *The Political Quarterly* 94 (2): 279-289. <https://doi.org/https://doi.org/10.1111/1467-923X.13267>.
- Borko, H. 1968. "Information Science: What Is It?" *American Documentation* 19 (1): 3-5. <https://doi.org/10.1002/asi.5090190103>.
- Bountouri, Lina, Matthew Damigos, Markella Drakiou, Manolis Gergatsoulis, and Eleftherios Kalogeros. 2023. "The semantic mapping of RiC-CM to CIDOC-CRM." *Leveraging Generative Intelligence in Digital Libraries: Towards Human-Machine Collaboration*:

- 25th International Conference on Asia-Pacific Digital Libraries, ICADL 2023, Taipei, Taiwan, December 4–7, 2023, Proceedings, Part II, 90–99. , Singapore, 2023//.
- Bowker, Geoffrey, and Susan Leigh Star. 2000. "Sorting things out : classification and Its consequences." *Boston, MA: MIT Press*.
- Bradley, John Douglas. 2020. "A Prosopography as Linked Open Data: Some Implications from DPRR." *DHQ: Digital Humanities Quarterly* 14 (2).
- Bradley, John, and Harold Short. 2005. "Texts into databases: the evolving field of new-style prosopography." *Literary and Linguistic Computing* 20 (Suppl): 3-24.
- Brennan, Claire. 2018. "Digital humanities, digital methods, digital history, and digital outputs: History writing and the digital revolution." *History Compass* 16 (10): e12492. <https://doi.org/https://doi.org/10.1111/hic3.12492>.
- Breure, Leen, Peter Doorn, and Onno Boonstra. 2006. *Past, present and future of historical information science*. The Hague: DANS.
- British, and Society Foreign Aborigines' Protection. 1844. *Seventh Annual Report of the Aborigines' Protection Society, presented at the meeting in Crosby Hall, May 20, 1844. With lists of officers, honorary and corresponding members, subscribers, and benefactors*.
- Brouwer, Judith, and Harm Nijboer. 2017. "Golden Agents. A Web of linked biographical data for the Dutch Golden Age." BD, In Proceedings of the Second Conference on Biographical Data in a Digital World 2017, Linz, Austria, November 6–7, 2017.
- Bruseker, George, Nicola Carboni, and Anaïs Guillem. 2017. "Cultural heritage data management: the role of formal ontology and CIDOC CRM." *Heritage and archaeology in the digital age: acquisition, curation, and dissemination of spatial cultural heritage data*. Edited by M. L. Vincent, V. M. López-Menchero Bendicho, M. Ioannides and T. E. Levy: 93-131.
- Byrum, John D. 2004. "NACO: A cooperative model for building and maintaining a shared Name Authority Database." *Authority Control in Organizing and Accessing Information: Definition and International Experience in Cataloging & Classification Quarterly* 38 (3-4): 237-249. https://doi.org/10.1300/J104v38n03_18.
- The Canadian Crisis and Lord Durham's Mission. 1838.
- Cannan, Judith P, Paul Frank, and Les Hawkins. 2019. "LC/NACO authority file in the library of congress BIBFRAME pilots." *Journal of Library Metadata* 19 (1-2): 39-51. <https://doi.org/https://doi.org/10.1080/19386389.2019.1589693>.
- Cantor, G. N. 2005. *Quakers, Jews, and science : religious responses to modernity and the sciences in Britain, 1650-1900*. Oxford: Oxford University Press.

- Careless, Virginia Ann Stockford. 1974. British Columbia.
- Carnegie, Garry D, and Karen M McBride. 2023. "Prosopography and microhistory: Illuminating historical actors." In *Handbook of Historical Methods for Management*, 245-263. Edward Elgar Publishing.
- Castells, Manuel. 2011. *The rise of the network society*. Chichester: John Wiley & Sons.
- Ceci, Michelangelo, Stefano Ferilli, and Antonella Poggi. 2020. *Digital libraries: The Era of big data and data science*. Cham: Springer.
- Champion, Erik Malcolm. 2017. "Digital humanities is text heavy, visualization light, and simulation poor." *Digital Scholarship in the Humanities* 32(Suppl 1): i25–i32: fqw053. <https://doi.org/10.1093/llc/fqw053>.
- Chave, Isabelle, and Claire Sibille-de Grimoüard. 2015. "Towards the development of a National Archival Authority File in France: an approach to Implement EAC-CPF." *Journal of Archival Organization* 12 (1-2): 98-117. <https://doi.org/https://doi.org/10.1080/15332748.2015.1000207>. .
- Ciro, Jennifer. 2002. "Country house libraries in the nineteenth-century." *Library History* 18 (2): 89-98.
- Clavaud, Florence. 2021. "Implementing ICA Records in Contexts-Ontology (RiC-O) at the National Archives of France (ANF): first steps and prospects." Le web sémantique et le patrimoine culturel: de la convergence des données au croisement des connaissances, Groupe d'Études et de Recherche Interdisciplinaire en Information et Communication (Gériico, Université de Lille), February 2021, Lille, France. .
- Cohen, Daniel J, Michael Frisch, Patrick Gallagher, Steven Mintz, Kirsten Sword, Amy Murrell Taylor, William G Thomas, and William J Turkel. 2008. "Interchange: the promise of digital history." *Journal of American History* 95 (2): 452-491.
- Crease, Robert, Elyse Graham, and Jamie Folsom. 2019. "Database thinking and deep description: designing a digital archive of the National Synchrotron Light Source." *Digital Scholarship in the Humanities* 34 (Supplement 1): i46-i57. <https://doi.org/https://doi.org/10.1093/llc/fqz053>.
- Crymble, Adam. 2021. *Technology and the historian: transformations in the digital age*. Vol. 1. Champaign, IL: University of Illinois Press.
- Dappert, Angela, and Markus Enders. 2010. "Digital preservation metadata standards." *Information Standards Quarterly* 22 (2): 4-13.
- de Groot, Jerome. 2015. "International Federation for Public History plenary address on genealogy." *The Public Historian* 37 (3): 102-127. <https://doi.org/https://doi.org/10.1525/tph.2015.37.3.102>. .

- . 2020. "Ancestry. com and the evolving nature of historical information companies." *The Public Historian* 42 (1): 8-28.
- Decker, Stefan, Sergey Melnik, Frank Van Harmelen, Dieter Fensel, Michel Klein, Jeen Broekstra, Michael Erdmann, and Ian Horrocks. 2000. "The semantic web: The roles of XML and RDF." *IEEE Internet Computing* 4 (5): 63-73.
- Delmas-Glass, Emmanuelle, and Robert Sanderson. 2020. "Fostering a community of PHAROS scholars through the adoption of open standards." *Art Libraries Journal* 45 (1): 19-23. <https://doi.org/https://doi.org/10.1017/alj.2019.32>.
- Duncan, Dennis. 2021. *Index, A history of the*. London: Allen Lane.
- Edmonds Penelope and Laidlaw Zoë. 2019. "'The British government Is now awaking': How humanitarian Quakers repackaged and circulated the 1837 Select Committee Report on aborigines." In *Aboriginal protection and its intermediaries in Britain's antipodean colonies*, edited by Samuel Furphy and Amanda Nettelbeck, 38-57. London: Routledge.
- Freund, Luanne, and Elaine G. Toms. 2016. "Interacting with archival finding aids." *Journal of the Association for Information Science and Technology* 67 (4): 994-1008. <https://doi.org/https://doi.org/10.1002/asi.23436>.
- Friendly, Michael. 2008a. "A brief history of data visualization." In *Handbook of Data Visualization*, edited by Wolfgang Härdle and Antony Unwin Chun-houh Chen, 15-56. Berlin: Springer.
- . 2008b. "The golden age of statistical graphics." *Statistical Science* 23 (4): 502-535. <https://doi.org/https://doi.org/10.1214/08-STS268>.
- Gamble, Clive. 2021. *Making deep history: zeal, perseverance, and the time revolution of 1859*. Oxford: Oxford University Press.
- Gavrilova, Natalia, and Leonid Gavrilov. 1999. "Data resources for biodemographic studies on familial clustering of human longevity." *Demographic Research* 1 (4). <https://doi.org/https://doi.org/10.4054/DemRes.1999.1.4>.
- Gilliland-Swetland Anne J., Philip B. Eppard. . 2000. "Preserving the authenticity of contingent digital objects. The iInterPARES project." *D-Lib Magazine* 6 (7/8). <https://www.dlib.org/dlib/july00/eppard/07eppard.html> (Accessed 22 March 2025).
- Goldenfein, Jake, and Daniel S Griffin. 2022. "Platforming the scholarly economy." *Internet Policy Review* 11 (3). <https://doi.org/https://doi.org/10.14763/2022.3.1671>.
- Gracy, Karen, and Frank Lambert. 2014. "Who's ready to surf the next wave? A study of perceived challenges to implementing new and revised standards for archival description." *The American Archivist* 77 (1): 96-132.

- Gusenbauer, Michael. 2019. "Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases." *Scientometrics* 118 (1): 177-214. <https://doi.org/10.1007/s11192-018-2958-5>.
<https://doi.org/10.1007/s11192-018-2958-5>.
- Gusenbauer, Michael, and Neal R Haddaway. 2020. "Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources." *Research synthesis Methods* 11 (2): 181-217.
- Haase, Peter, Daniel M Herzig, Artem Kozlov, Andriy Nikolov, and Johannes Trame. 2019. "metaphactory: a platform for knowledge graph management." *Semantic Web* 10 (6): 1109-1125.
- Haggerty, John, and Sheryllynne Haggerty. 2011. "The life cycle of a metropolitan business network: Liverpool 1750–1810." *Explorations in Economic History* 48 (2): 189-206. <https://doi.org/https://doi.org/10.1016/j.eeh.2010.09.006>.
- Haggerty, Sheryllynne. 2006. "The British-Atlantic Trading Community, 1760–1810." *Women, and the Distribution of Goods*. Brill: Leiden-Boston.
- . 2008. "Liverpool, the slave trade and the British-Atlantic empire, c. 1750–75." In *The empire in one city?*, 17-34. Manchester University Press.
- Hamm, Thomas D. 2013. "Hicksite, Orthodox, and Evangelical Quakerism, 1805–1887." In *Oxford Handbook of Quaker studies*, edited by Stephen W. Angell and Pink Dandelion, 63 - 77. Oxford: Oxford University Press.
- Hammond, Matthew. 2021. "From digital prosopography to social network analysis." *Medieval People: Social Bonds, Kinship, and Networks* 36: 235-262.
- Heartfield, James. 2011. *The Aborigines' Protection Society : humanitarian imperialism in Australia, New Zealand, Fiji, Canada, South Africa, and the Congo, 1837-1909*. London: Hurst.
- Hering, Katharina 2014. "Provenance meets source criticism." *Journal of Digital Humanities* 3, no. 2. <https://journalofdigitalhumanities.org/3-2/provenance-meets-source-criticism> (Accessed 22 March 2025).
- Hering, Katharina , Michael J. Kramer, Joshua Sternfeld, and Kate Theimer. 2014. "Digital Historiography and the Archives." *Journal of Digital History* 3, no. 2. <https://journalofdigitalhumanities.org/3-2/digital-historiography-and-the-archives> (Accessed 22 March 2025).
- Higgs Edward, Jones Christine, Schürer Kevin, and Wilkinson Amanda. 2021. *I-CeM Guide*<https://www.essex.ac.uk/research-projects/integrated-census-microdata>.

- Higgs, Edward, Christine Jones, Kevin Schürer, and Amanda Wilkinson. 2021. *Integrated Census Microdata (I-CeM) guide*. <https://www.essex.ac.uk/research-projects/integrated-census-microdata> (Accessed 22 March 2025).
- Hilbert, M. 2020. Digital technology and social change: The digital transformation of society from a historical perspective. *Dialogues in Clinical Neuroscience* 22(2): 189-194. <https://doi.org/https://doi.org/10.31887/DCNS.2020.22.2/mhilbert>.
- Hilbert, Martin, and Priscila López. 2011. "The World's technological capacity to store, communicate, and compute Information." *Science* 332 (6025): 60-65. <https://doi.org/https://doi.org/doi:10.1126/science.1200970>.
- Hill, R.A., and R.I.M Dunbar. 2003. "Social network size in humans." *Human Nature* 14 (1): 53-72. <https://doi.org/https://doi.org/10.1007/s12110-003-1016-y>.
- Hjorthén, Adam. 2022. "An ocean of information: labour, commodification, and the culture of indexes in modern transatlantic genealogy." *Library & Information History* 38 (3): 189-209.
- Hockey, Susan. 2004. *The history of humanities computing*. Edited by Ray Siemens and John Unsworth Susan Schriebman. *A companion to digital humanities*. Oxford: Blackwell.
- Hodgkin, Thomas. 1837. Minutes of first meeting of the Aborigines Protection Society. Wellcome Library; GB.
- . 1848. "The Progress of Ethnology." *Journal of the Ethnological Society of London (1848-1856)* 1: 27-45. <https://doi.org/https://doi.org/10.2307/3014076>. . <http://www.jstor.org/stable/3014076>.
- Hoeve, Casey Daniel. 2018. "Finding a place for genealogy and family history in the digital humanities." *Digital Library Perspectives* 34 (3): 215-226.
- Hyvönen, Eero, Petri Leskinen, Minna Tamper, Heikki Rantala, Esko Ikkala, Jouni Tuominen, and Kirsi Keravuori. 2019. "Demonstrating BiographySampo in solving digital humanities research problems in biography and prosopography." *Digital Humanities in the Nordic Countries*.
- Idrissou, Al, Leon Van Wissen, and Veruska Zamborlini. 2022. "The Lenticular Lens: Addressing various aspects of entity disambiguation in the Semantic Web; 2022." *Graphs and Networks in the Humanities*: 3-4. <https://www.leonvanwissen.nl/publication/idrissouetal-2022-graphum> (Accessed 22 March 2025).
- Idrissou, Al, Veruska Zamborlini, Chiara Latronicoc, Frank van Harmelenvan, and Charles den Heuvel. 2018. *Amsterdammers from the Golden Age to the Information Age via Lenticular Lenses*. Amsterdam,: DHBenelux.

<https://pure.know.nl/portal/en/publications/amsterdammers-from-the-golden-age-to-the-information-age-via-lenti> (Accessed 22 March 2025).

International Council on Archives. 1999. *ISAD (G) : general international standard archival description*. ICA (Washington D.C.).

---. 2004. *International Standard Archival Authority Record for Corporate Bodies, Persons and Families*, ed International Council on Archives. International Standard. Paris: ICA. https://www.ica.org/app/uploads/2023/12/CBPS_Guidelines_ISAAR_Second-edition_EN.pdf (Accessed 22 March 2025).

---. 2021. *Archival arrangement & description: global practices*. ICA (Paris). https://www.ica.org/app/uploads/2023/12/aad_survey_report_final_202108_eng.pdf (Accessed 22 March 2025).

---. 2023. *Records in Contexts Conceptual Model*. . Expert Group on Archival Description, INTERNATIONAL COUNCIL ON ARCHIVES (Paris: ICA). <https://www.ica.org/resource/records-in-contexts-conceptual-model/> (Accessed 22 March 2025).

International Federation of Library Associations and Institutions. 2009. *Functional requirements for bibliographic records, final report*. International Federation of Library Associations and Institutions (The Hague, IFLA). <https://repository.ifla.org/handle/123456789/811> (Accessed 22 March 2025).

Jockers, Matthew L. 2013. *Macroanalysis: digital methods and literary history*. Champaign IL: University of Illinois Press.

Jones, Steven E. 2013. *The emergence of the digital humanities*. New York: Taylor & Francis.

Jordanova, Ludmilla. 2019. *History in practice*. London: Bloomsbury.

Kaltenbrunner, Wolfgang. 2017. "Digital Infrastructure for the humanities in Europe and the US: governing scholarship through coordinated tool development." *Computer Supported Cooperative Work (CSCW)* 26 (3): 275-308. <https://doi.org/https://doi.org/10.1007/s10606-017-9272-2>.

Kansa, Sarah Witcher, and Eric C Kansa. 2018. "Data beyond the archive in digital archaeology: an introduction to the special section." *Advances in Archaeological Practice* 6 (2): 89-92.

Kass, Amalie M., and Kass, Edward H. 1988. *Perfecting the world : the life and times of Dr. Thomas Hodgkin 1798-1866* Edited by Edward H. Kass. 1st ed. Boston: Harcourt Brace Jovanovich.

Keats-Rohan, Katharine SB. 2007. "Biography, identity and names: understanding the pursuit of the individual in prosopography." *Prosopography Approaches and*

- Applications. A Handbook*, edited by Katharine S. B. Keats-Rohan, 139-182. Oxford: Oxford University Press.
- Kellert, Stephen H. 2008. "The rhetorical functions of borrowing and the uses of disciplinary prestige." In *Borrowed Knowledge: chaos theory and the challenge of learning across disciplines*, 57-80. Chicago IL: University of Chicago Press.
- Kennard, Douglas J, Andrew M Kent, and William A Barrett. 2011. "Linking the past: discovering historical social networks from documents and linking to a genealogical database." Proceedings of the 2011 workshop on historical document Imaging and processing.
- Kennard, Douglas J, William B Lund, and Bryan S Morse. 2009. "Improving historical research by linking digital library information to a global genealogical database." Proceedings of the 9th ACM/IEEE-CS joint conference on digital libraries.
- Kirschenbaum, Matthew. 2014. "What is "Digital Humanities," and why are they saying such terrible things about it?" *differences* 25 (1): 46-63.
- Kirschenbaum, Matthew, Richard Ovenden, Gabriela Redwine, and Rachel Donahue. 2010. *Digital forensics and born-digital content in cultural heritage collections*. Alexandria, VA: Council on Library and Information Resources.
<https://www.clir.org/pubs/reports/pub149> (Accessed 22 March 2025).
- Koolen, Marijn, Jasmijn Van Gorp, and Jacco Van Ossenbruggen. 2019. "Toward a model for digital tool criticism: reflection as integrative practice." *Digital Scholarship in the Humanities* 34 (2): 368-385. <https://doi.org/https://doi.org/10.1093/llc/fqy048>. .
- Kowaleski, Maryanne. 2021. "A New Digital Prosopography: The Medieval Londoners Database." *Medieval People* 36 (1): 13.
- Laidlaw, Zoë. 2001. "Networks, patronage and information in colonial governance : Britain, New South Wales and the Cape Colony, 1826-1843." PhD, University of Oxford.
- . 2004. "'Aunt Anna's report' : the Buxton women and the Aborigines Select Committee, 1835-37." *Journal of Imperial and Commonwealth History* 32(2): 1-28.
- . 2005a. *Colonial connections, 1815-45 : patronage, the information revolution and colonial government*. Manchester: Manchester University Press.
- . 2005b. *Colonial connections, 1815-45 : patronage, the information revolution and colonial government* / Zoë Laidlaw. Manchester: Manchester University Press.
- . 2007. "Heathens, slaves and aborigines: Thomas Hodgkin's critique of missions and anti-slavery." *History Workshop Journal* (64): 133-161.
<https://doi.org/https://doi.org/10.1093/hwj/dbm034>.

- . 2021. *Protecting the Empire's Humanity: Thomas Hodgkin and British Colonial Activism 1830–1870. Critical Perspectives on Empire*. Cambridge: Cambridge University Press.
- Lampron, Patricia, and Melanie Wacker. 2019. "Name authority work in the linked data environment." *Journal of Library Metadata* 19 (1-2): 137-140.
<https://doi.org/https://doi.org/10.1080/19386389.2019.1661109>.
- Laslett, Peter, and Kevin Schürer. 2021. *The world we have lost*. London: Routledge.
- Lemercier, Claire. 2015. "Formal network methods in history: why and how?" In *Social networks, political institutions, and rural societies*, 281-310. Turnhout: Brepols.
- Lester, Alan. 2005. *Imperial networks: creating identities in nineteenth-century South Africa and Britain*. London: Routledge.
- Lester, Alan, and Fae Dussart. 2014. *Colonization and the origins of humanitarian governance : protecting aborigines across the nineteenth-century British Empire*. Cambridge: Cambridge University Press.
- Library of Congress. 2009. "Understanding MARC Bibliographic: Machine-Readable Cataloging." Network Development and MARC Standards Office, Library of Congress. Accessed 04/12/2023. <https://www.loc.gov/marc/umb/>.
- . 2019. Library of Congress BIBFRAME manual.
- Library of the Society of Friends. 1831. London Yearly Meeting. Friends House Archive, Quakers in Britain, London.
- . 1832. London Yearly Meeting Friends House Archive, Quakers in Britain, London.
- . 1833. London Yearly Meeting. Friends House Archive, Quakers in Britain, London.
- . 1834. London Yearly Meeting Friends House Archive, Quakers in Britain, London.
- . 1835. London Yearly Meeting. Friends House Archive, Quakers in Britain, London.
- . 1837a. London Yearly Meeting. Friends House Archive, Quakers in Britain, London.
- . 1837b. London Yearly Meeting Meetings for Sufferings Friends House Archive, Quakers in Britain, London.
- . 1838a. London Yearly Meeting. Friends House Archive, Quakers in Britain, London.
- . 1838b. London Yearly Meeting Meetings for Sufferings Friends House Archive, Quakers in Britain, London.
- . 1839. London Yearly Meeting Meetings for Sufferings Friends House Archive, Quakers in Britain, London.
- . 1840a. London Yearly Meeting. Friends House Archive, Quakers in Britain, London.

- . 1840b. London Yearly Meeting Meetings for Sufferings. Friends House Archive, Quakers in Britain, London.
- . 1842. London Yearly Meeting Meetings for Sufferings Friends House Archive, Quakers in Britain, London.
- . 1844. London Yearly Meeting Meetings for Sufferings. Friends House Archive, Quakers in Britain, London.
- . 1845. London Yearly Meeting Meetings for Sufferings Friends House Archive, Quakers in Britain, London.
- . 1846a. London Yearly Meeting. Friends House Archive, Quakers in Britain, London.
- . 1846b. London Yearly Meeting Meetings for Sufferings. Friends House Archive, Quakers in Britain, London.

Library of the Society of Friends. LYM Meeting for Sufferings. 1837. *Information Respecting the Aborigines in the British Colonies*. In *BEING PRINCIPALLY EXTRACTS FROM THE REPORT PRESENTED TO THE HOUSE OF COMMONS, BY THE SELECT COMMITTEE APPOINTED ON THAT SUBJECT*. London: Darton and Harvey.

https://archive.org/details/cihm_21680 (Accessed 22 March 2025).

- . 1838. *Information respecting the aborigines in the British colonies circulated by direction of the Meeting for Sufferings : being principally extracts from the report presented to the House of Commons, by the select committee appointed on that subject*. London:

Darton and Harvey. https://archive.org/details/cihm_21680

https://archive.org/details/cihm_21680 (Accessed 22 March 2025).

- . 1839a. *Facts relative to the Canadian Indians published by direction of the Aborigines' Committee of the Meeting for Sufferings*. London: Darton and Harvey.

https://archive.org/details/cihm_21758 (Accessed 22 March 2025).

- . 1839b. *Further information respecting the aborigines containing extracts from the proceedings of the Meeting for Sufferings in London, and of the Committees on Indian Affairs of the yearly meetings of Philadelphia and Baltimore; together with some particulars relative to the Seminole War*. Microform. London: Darton and Harvey.

https://archive.org/details/cihm_42219

https://archive.org/details/cihm_21680 (Accessed 22 March 2025).

- . 1840. *The report of the Aborigines' Search Committee of the Meeting for Sufferings, read at the yearly meeting 1840 with the address to Lord John Russell on his becoming secretary for the colonies; that to Friends settling in new colonies, and some particulars calculated to give information and promote interest respecting the present state of aboriginal tribes*. London: Darton and Harvey.

- https://archive.org/details/cihm_41787 https://archive.org/details/cihm_21680 (Accessed 22 March 2025).
- . 1841. *The report of the Meeting for Sufferings respecting the aborigines, presented to the Yearly Meeting, 1841* London: Darton and Harvey.
https://archive.org/details/cihm_18635 https://archive.org/details/cihm_21680 (Accessed 22 March 2025).
- . 1842. *Further information respecting the aborigines containing reports of the committee on Indian affairs at Philadelphia, extracts from the proceedings of the yearly meetings of Philadelphia, New York, New England, Maryland, Virginia, and Ohio, together with some particulars relative to the natives of New Zealand, New Holland, and Van Dieman's land.* London: Edward Marsh.
https://archive.org/details/cihm_41786 https://archive.org/details/cihm_21680 (Accessed 22 March 2025).
- . 1843. *Further information respecting the aborigines containing extracts from the proceedings of the Meeting for Sufferings in London, and of the Committees on Indian Affairs, of the yearly meetings of Philadelphia and Baltimore; together with some particulars relative to the Seminole War.* London: Edward Marsh.
https://archive.org/details/cihm_49768 https://archive.org/details/cihm_21680 (Accessed 22 March 2025).
- . 1844. *Some account of the conduct of the religious Society of Friends towards the Indian tribes in the settlement of the colonies of east and west Jersey and Pennsylvania with a brief narrative of their labours for the civilization and Christian instruction of the Indians, from the time of their settlement in America, to the year 1843.* London: Edward Marsh. https://archive.org/details/cihm_40285 https://archive.org/details/cihm_21680 (Accessed 22 March 2025).
- Liu, Alan. 2013. "The Meaning of the digital humanities." *PMLA* 128 (2): 409-423.
<https://doi.org/https://doi.org/10.1632/pmla.2013.128.2.409>.
- Llanes-Padrón, Dunia, and Juan-Antonio Pastor-Sánchez. 2017. "Records in contexts: the road of archives to semantic interoperability." *Program* 51 (4): 387-405.
<https://doi.org/https://doi.org/10.1108/PROG-03-2017-0021>.
- Lubenow, William C. 2015. *"Only connect". Learned societies in nineteenth-century Britain.* Woodbridge: Boydell & Brewer.
- Mahon, Basil. 2009. Knowledge is Power. A short history of official data collection in the UK.
- Marchese, Francis T. 2011. "Exploring the origins of tables for information visualization." 15th International Conference on Information Visualisation.

- Martín-Martín, Alberto, Mike Thelwall, Enrique Orduna-Malea, and Emilio Delgado López-Cózar. 2021. "Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations." *Scientometrics* 126 (1): 871-906.
<https://doi.org/https://doi.org/10.1007/s11192-020-03690-4>.
- Maxwell, Robert L. 2001. *Maxwell's guide to authority work*. Chicago: ALA Editions.
- Mayernik, Matthew S. 2019. "Metadata accounts: achieving data and evidence in scientific research." *Social Studies of Science* 49 (5): 732-757.
- . 2021. "Metadata." *KO Knowledge Organization* 47 (8): 696-713.
- Meroño-Peñuela, Albert, Ashkan Ashkpour, Marieke Van Erp, Kees Mandemakers, Leen Breure, Andrea Scharnhorst, Stefan Schlobach, and Frank Van Harmelen. 2015. "Semantic technologies for historical research: a survey." *Semantic Web* 6 (6): 539-564.
- Millar, Laura. 2002. "The death of the fonds and the resurrection of provenance: archival context in space and time." *Archivaria* 53: 1-15.
- Miller, Randy L. 2019. "An Introduction to Google Scholar."
- Milligan, Edward. 2009. *Biographical dictionary of British Quakers in commerce and industry, 1775-1920*. York: Sessions Book Trust.
- Milligan, Ian. 2022. *The transformation of historical research in the digital age*. Cambridge: Cambridge University Press.
- Million, Anthony J., Jeremy York, Sara Lafia, and Libby Hemphill. 2024. "Data, not documents: Moving beyond theories of information-seeking behavior to advance data discovery." *Journal of the Association for Information Science and Technology*.
<https://doi.org/https://doi.org/10.1002/asi.24962>.
- Morgan, Colleen Leah. 2012. *Emancipatory digital archaeology*. Berkeley CA: University of California, .
- Morris, Sammie L. 2009. "An introduction to archives for librarians." *Libraries Research Publications* 103. http://docs.lib.purdue.edu/lib_research/103 (Accessed 22 March 2025).
- Morrish, P. S. 2006. "Library management in the pre-professional age." In *The Cambridge History of Libraries in Britain and Ireland: Vol 2: 1640–1850*, edited by Giles Mandelbrote and K. A. Manley, 479-493. Cambridge: Cambridge University Press.
- Munslow, Alun. 1997. "Review of What is History?" *Rethinking History: The Journal of Theory and Practice*, 41a.
<https://doi.org/https://doi.org/10.14296/RiH/issn.1749.8155>. .

- Murtomaa, Eeva. 2000. "The impact of the functional requirements for bibliographic records recommendations on the ISBD (ER)." *Cataloging & Classification Quarterly* 28 (1): 33-41.
- Nash, Kate. 2001. "The 'cultural turn' in social theory: towards a theory of cultural politics." *Sociology* 35 (1): 77-92.
<https://doi.org/https://doi.org/10.1017/S0038038501000050>.
- Nyhan, Julianne, and Andrew Flinn. 2016. *Computation and the humanities: towards an oral history of digital humanities*. Cham: Springer Nature.
- O'Hare, Sheila. 2002. "Genealogy and History." *Common Place* 2 (3).
<https://commonplace.online/article/genealogy-and-history/?print=pdf> (Accessed 22 March 2025).
- Oldman, Dominic. 2014. "The CIDOC conceptual reference model (CIDOC-CRM): primer." *CRM Labs* 5. https://cidoc-crm.org/sites/default/files/CRMPrimer_v1.1_1.pdf (Accessed 22 March 2025).
- Oldman, Dominic, Martin Doerr, and Stefan Gradmann. 2015. "Zen and the art of linked data: new strategies for a semantic web of humanist knowledge." In *A new companion to digital humanities*, edited by Ray Siemens and John Unsworth Susan Schreibman, 251-273. Chichester: John Wiley and Sons.
- Oldman, Dominic, and Diana Tanase. 2018. "Reshaping the knowledge graph by connecting researchers, data and practices in ResearchSpace." *The Semantic Web – ISWC 2018: 17th international semantic web conference* Monterey CA.
- Oldman, Dominic, Diana Tanase, and Stephanie Santschi. 2019. "The problem of distance in digital art history: a ResearchSpace case study on sequencing Hokusai print impressions to form a human curated network of knowledge." *International Journal for Digital Art History* (4): 5.29-5.45.
<https://doi.org/https://doi.org/10.11588/dah.2019.4.72071>.
- Oliver, Chris. 2021. *Introducing RDA : a guide to the basics after 3R*. Chicago, IL: American Library Association.
- Olivier, Philippe, Martin Hammitzsch, Stephan Janosch, Anelda van der Walt, Ben van Werkhoven, Simon Hettrick, Daniel S. Katz, Katrin Leinweber, Sandra Gesing, Stephan Druskat, Scott Henwood, Nicholas R. May, Nooriyah P. Lohani, and Manodeep Sinha. 2019. Shaping data and software policy in the arts and humanities reserch community. <https://doi.org/https://doi.org/10.5281/zenodo.2585783>.
- Online Computer Library Center. 2019. "VIAF guidelines." OCLC.
<https://www.oclc.org/en/viaf/contributing.html> (Accessed 22 March 2025).

- Orduna-Malea, Enrique, Juan M. Ayllón, Alberto Martín-Martín, and Emilio Delgado López-Cózar. 2015. "Methods for estimating the size of Google Scholar." *Scientometrics* 104 (3): 931-949. <https://doi.org/https://doi.org/10.1007/s11192-015-1614-6>.
- Otlet, Paul, and W Boyd Rayward. 1990. "International organisation and dissemination of knowledge: selected essays of Paul Otlet." *FID*; 684.
- Pacheco, André, Carlos Guardado Da Silva, and Maria Cristina Vieira De Freitas. 2023. "A metadata model for authenticity in digital archival descriptions." *Archival Science* 23 (4): 629-673. <https://doi.org/https://doi.org/10.1007/s10502-023-09422-w>.
- Papadakis, Ioannis, Konstantinos Kyprianos, and Michalis Stefanidakis. 2015. "Linked data URIs and libraries: the story so far." *D-Lib Magazine* 21 (5-6): 1. <https://doi.org/https://doi.org/10.1045/may2015-papadakis>.
- Papers, APS. 1839. "<Extracts from The Papers and Proceedings of the Aborigines Protection Society 1837.pdf>."
- Pawlicka-Deger, Urszula. 2021. "Infrastructuring digital humanities: on relational infrastructure and global reconfiguration of the field." *Digital Scholarship in the Humanities* 37 (2): 534-550. <https://doi.org/https://doi.org/10.1093/llc/fgab086>.
- Piersma, Hinke, and Kees Ribbens. 2013. "Digital historical research: context, concepts and the need for reflection." *BMGN-Low Countries Historical Review* 128 (4): 78-102.
- Piotrowski, Michael, and Mateusz Fafinski. 2020. "Nothing New Under the Sun? Computational humanities and the methodology of history." CHR 2020: Workshop on computational humanities research., Amsterdam.
- Pitti, Daniel, Rachael Hu, Ray Larson, Brian Tingle, and Adrian Turner. 2015. "Social networks and archival context: From project to cooperative archival program." *Journal of Archival Organization* 12 (1-2): 77-97. <https://doi.org/https://doi.org/10.1080/15332748.2015.999544>.
- Rainger, Ronald. 1976. "The organizational development anthropology in England 1837-1871." MA, University of Utah, Jisc.
- . 1980. "Philanthropy and science in the 1830's: The British and Foreign Aborigines' Protection Society." *Man* 15 (4): 702-717. <https://doi.org/https://doi.org/10.2307/2801541>. .
- Ramdurai, Balagopal, and Prasanna Adhithya. 2023. "The impact, advancements and applications of generative AI." *International Journal of Computer Science and Engineering* 10 (6): 1-8.
- Ramsay, Stephen. 2004. "Databases." In *A companion to digital humanities*, edited by Ray Siemens and John Unsworth Susan Schriebman, 177-197. Oxford: Blackwell.

- Ranade, Sonia. 2016. "Traces through time: A probabilistic approach to connected archival data." 2016 IEEE International Conference on Big Data (Big Data), Washington, DC.
- Reynaert, Martin, Patrick Bos, and Janneke van der Zwaan. 2019. "Granularity versus dispersion in the Dutch diachronical database of Lexical frequencies TICCLAT." CLARIN Annual Conference
- Riley, Jenn. 2017. Understanding metadata. What is metadata, and what is it for? A primer. 23: 7-10.
- Rocha, Antonio Penalves. 2009. *Abolicionistas brasileiros e ingleses : a coligação entre Joaquim Nabuco e a British and Foreign Anti-slavery Society (1880-1902)*. São Paulo: Editora UNESP : Brazilian Business School.
- Romein, C Annemieke, Max Kemman, Julie M Birkholz, James Baker, Michel De Gruijter, ALBERT MEROÑO-PEÑUELA, Thorsten Ries, Ruben Ros, and Stefania Scagliola. 2020. "State of the field: digital history." *History* 105 (365): 291-312.
- Rosenfeld, Louis. 2000. "Thomas Hodgkin: social activist." *Annals of Diagnostic Pathology* 4 (2): 124-133.
- Roueché, Charlotte, Averil Cameron, and Janet L Nelson. 2023. "Prosopography meets the digital: PBW and PASE." *On Making in the Digital Humanities. The Scholarship of Digital Humanities Development in Honour of John Bradley*: 51-65.
- Ryan, Marie-Laure, Lori Emerson, and Benjamin J Robertson. 2014. *The Johns Hopkins guide to digital media*. Baltimore MD: John's Hopkins University Press.
- Sætra, Henrik Skaug. 2023. "Generative AI: Here to stay, but for good?" *Technology in Society* 75: 102372. <https://doi.org/https://doi.org/10.1016/j.techsoc.2023.102372>.
- Scheinfeldt, Tom. 2008. "Sunset for ideology, sunrise for methodology? ." *Found History* (30/10/2023). <http://foundhistory.org/2008/03/sunset-for-ideology-sunrise-for-methodology/> (Accessed 30 October 2023).
- Schöch, Christof. 2013. "Big? Smart? Clean? Messy? Data in the humanities?" *Journal of the Digital Humanities* 2 (3). <https://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities> (Accessed 22 March 2025).
- Schreibman, Susan, Ray Siemens, and John Unsworth. 2015. *A new companion to digital humanities*. Chichester: John Wiley & Sons.
- Schwandt, Silke. 2022. "Opening the black box of interpretation: digital history practices as models of knowledge." *History and Theory* 61 (4): 77-85. <https://doi.org/https://doi.org/10.1111/hith.12281>.
- Seefeldt, Douglas, and William G Thomas III. 2009. What is digital history? A look at some exemplar projects. *Faculty Publications, Department of History*. .

- Sera-Shriar, Efram. 2013. *The making of British anthropology, 1813-1871*. Pittsburgh PA: University of Pittsburgh Press.
- Siebold, Anna, and Matteo Valleriani. 2022. "Digital perspectives in history." *Histories* 2 (2): 170-177. <https://doi.org/https://doi.org/10.3390/histories2020013>.
- Smithies, James. 2017. *The digital humanities and the digital modern*. Cham: Springer.
- Society of American Archivists. 2020. Describing Archives: A Content Standard - DACS. (Version 2019.0.3.). Accessed 2019.0.3. <https://doi.org/ISBN: 978-1-945246-97-5>.
- . 2023. Encoded Archival Description Tag Library Version EAD3 1.1.2,. <https://doi.org/978-1-958954-15-7>.
- Spina, Salvatore. 2021. "The digital age of historians." *AIDAinformazioni*: 103-120. <https://doi.org/https://doi.org/10.57574/5965090006>.
- Star, Susan, and Karen Ruhleder. 1996. "Steps toward an ecology of infrastructure: design and access for large information spaces." *Information Systems Research* 7: 111-134. <https://doi.org/https://doi.org/10.1287/isre.7.1.111>.
- Stead, Stephen. 2008. "The CIDOC CRM, a standard for the integration of cultural information" ((Powerpoint presentation)). <https://www.cidoc-crm.org/cidoc-crm-tutorial> (Accessed 22 March 2025).
- Sternfeld, Joshua. 2014. "Historical understanding in the quantum age." *Journal of Digital Humanities* 3, no. 2. <https://journalofdigitalhumanities.org/3-2/digital-contexts/> (Accessed 22 March 2025).
- Stichweh, Rudolf. 1992. "The Sociology of Scientific Disciplines: On the Genesis and Stability of the Disciplinary Structure of Modern Science." *Science in Context* 5 (1): 3-15. <https://doi.org/10.1017/S0269889700001071>. <https://www.cambridge.org/core/product/4D1D770C993ADFD8748E814A108F763>.
- Stocking, George W. 1971. "What's in a name? The origins of the Royal Anthropological Institute (1837-71)." *Man* 6 (3): 369-390. <https://doi.org/https://doi.org/10.2307/2799027>.
- . 1987. *Victorian Anthropology*. New York Free Press.
- Stubenrauch, Joseph. 2016. *The evangelical age of ingenuity in industrial Britain*. Oxford: Oxford University Press.
- Sufferings, Meetings for. 1838. *Information Respecting the Aborigines in British Colonies*. London: Dalton and Harvey. (accessed 1838).
- . 1839. *Further information respecting the aborigines [microform] : containing extracts from the proceedings of the Meeting for Sufferings in London, and of the Committees*

on Indian Affairs of the yearly meetings of Philadelphia and Baltimore; together with some particulars relative to the Seminole War.

Microform https://archive.org/details/cihm_42219.

- Sula, Chris Alen, and Heather V Hill. 2019. "The early history of digital humanities: An analysis of Computers and the Humanities (1966–2004) and Literary and Linguistic Computing (1986–2004)." *Digital Scholarship in the Humanities* 34, no. (Suppl 1: 190-206. <https://doi.org/https://doi.org/10.1093/lc/fqz072>.
- Svensson, Patrik. 2016. *Big digital humanities: imagining a meeting place for the humanities and the digital*. Ann Arbor, MI: University of Michigan Press.
- TANAKA, STEFAN. 2022. "THE OLD AND NEW OF DIGITAL HISTORY." *History and Theory* 61 (4): 3-18. <https://doi.org/https://doi.org/10.1111/hith.12284>.
- Terras, Melissa. 2016. "Peering inside the big tent: digital humanities and the crisis of inclusion." In *Defining Digital Humanities*, 263-270. London: Routledge.
- Terras, Melissa, James Baker, James Hetherington, David Beavan, Martin Zaltz Austwick, Anne Welsh, Helen O'Neill, Will Finley, Oliver Duke-Williams, and Adam Farquhar. 2018. "Enabling complex analysis of large-scale digital collections: humanities research, high-performance computing, and transforming access to British Library digital collections." *Digital Scholarship in the Humanities* 33 (2): 456-466. <https://doi.org/https://dx.doi.org/10.1093/lc/fqx020>.
- Thaller, Manfred. 2012. "Controversies around the digital humanities: an agenda." *Historical Social Research* 37-3: 7-23.
- Thomas, William G. 2004. "Computing and the historical imagination." In *A companion to digital humanities*, edited by Ray Siemens and John Unsworth Susan Schriebman, 56-68. Oxford: Blackwell.
- Tian, Cindy Tang, Timothy W Cole, and Karen Yu. 2021. "Name and subject heading reconciliation to linked open data authorities using virtual international authority file and library of congress linked data service APIs: a case study featuring emblematica online." *Library Resources & Technical Services* 65 (4): 132-142. <https://doi.org/http://dx.doi.org/10.2139/ssrn.5059848>.
- Tilly, Charles. 1990. "How (and what) are historians doing?" *American Behavioral Scientist* 33 (6): 685-711. <https://doi.org/https://doi.org/10.1177/0002764290033006005>.
- Tosh, John, and Sean Lang. 2021. *The pursuit of history: Aims, methods and new directions in the study of modern history*. 7th ed. London: Pearson Education.
- Twomey, Christina. 2018. "Protecting slaves and aborigines." *Pacific Historical Review* 87 (1): 10-29. <https://doi.org/https://doi.org/10.1525/phr.2018.87.1.10>.

- Unsworth, John. 2000. "Scholarly primitives: what methods do humanities researchers have in common, and how might our tools reflect this." Symposium on humanities computing: formal methods, experimental practice. , King's College, London.
- . 2002. "What is humanities computing and what is not?" *Jahrbuch für Computerphilologie* 4: 71-83. <http://hdl.handle.net/2142/157> (Accessed 22 March 2025).
- van Wissen, Leon, Chiara Latronico, Veruska Zamborlini, Jirsi Reinders, and CMJM van den Heuvel. 2020. "Unlocking the archives. A pipeline for scanning, transcribing and modelling entities of archival documents into linked open data." DH Benelux 2020.
- van Wissen, Leon, Veruska Zamborlini, and Charles van den Heuvel. 2022. "Modeling provenance and uncertainties in the use of archival sources of the Dutch Golden Age." The 68th annual meeting of the Renaissance Society of America, Dublin, Ireland.
- van Zundert, Joris. 2012. "If you build It, will we come? large scale digital infrastructures as a dead end for digital humanities." *Historical Social Research / Historische Sozialforschung* 37 (3): 165-186. <https://doi.org/https://doi.org/10.12759/hsr.37.2012.3.165-186>.
- Velios, Athanasios. 2019. "Linked Conservation Data / Ligatus terminology report (phase 1)." University of the Arts London. <https://www.ligatus.org.uk/lcd/output/142> (Accessed 22 March 2025).
- Velios Athanasios and St. John Kristen 2021. *Linked Conservation Data - Board reattachment pilot*. <https://www.ligatus.org.uk/lcd/>.
- Verboven, Koenraad, Myriam Carlier, and Jan Dumolyn. 2007. "A short manual to the art of prosopography." In *Prosopography approaches and applications. A handbook*, 35-70. Oxford: Unit for Prosopographical Research (Linacre College).
- Wagner, H Daniel. 2006. "Genealogy as an academic discipline." *AVOTAYNU* 22 (1): 6.
- Warwick, C., M. Terras, P. Huntington, and N. Pappa. 2007. "If you build it will they come? The LAIRAH study: quantifying the use of online resources in the arts and humanities through statistical analysis of user log data." *Literary and Linguistic Computing* 23 (1): 85-102. <https://doi.org/https://dx.doi.org/10.1093/lc/fqm045>.
- Warwick, Claire. 2015. "Building theories or theories of building? A tension at the heart of digital humanities." In *A new companion to digital humanities*, edited by Ray Siemens and John Unsworth Susan Schriebman, 538-552. Oxford: Blackwell.
- Waters, Donald J. 2023. "The emerging digital infrastructure for research in the humanities." *International Journal on Digital Libraries* 24 (2): 87-102. <https://doi.org/https://doi.org/10.1007/s00799-022-00332-3>.

- Weil, François. 2007. "John Farmer and the making of American genealogy." *New England Quarterly* 80 (3): 408-434.
<https://doi.org/https://doi.org/10.1162/tneq.2007.80.3.408>.
- Wiedeman, Gregory. 2019. "The historical hazards of finding aids." *The American Archivist* 82 (2): 381-420. <https://doi.org/https://doi.org/10.17723/aarc-82-02-20>.
- Willer, Mirna, and Gordon Dunsire. 2013. *Bibliographic information organization in the semantic web*. 1 ed. San Diego CA: Elsevier Science.
- Wilson, Patrick. 1983. "The catalog as access mechanism: background and concepts." *Library Resources and Technical Services* 27 (1): 4-17.
- Yakel, Elizabeth, and Deborah Torres. 2007. "Genealogists as a community of records". *The American Archivist* 70 (1): 93-113.
- Yaman, Beyza, Lucy McKenna, Alex Randles, Lynn Kilgallon, Peter Crooks, and Declan O'Sullivan. 2024. "Digital Prosopography Information in Virtual Record Treasury of Ireland Knowledge Graph." Proceedings of the 1st International Workshop of Semantic Digital Humanities (SemDH) Co-Located with the 21st Extended Semantic Web Conference.
- Zaagsma, Gerben. 2013. "On digital history." *BMGN-Low Countries Historical Review* 128 (4): 3-29.
- Zandhuis, Ivo. 2005. "Towards a genealogical ontology for the semantic web." Humanities, computers and cultural heritage: Proceedings of the XVI international conference of the Association for History and Computing.

