DHQ: Digital Humanities Quarterly
2009
Volume 3 Number 4

The Potential and Problems in using High Performance Computing in the Arts and Humanities: the Researching e-Science Analysis of Census Holdings (ReACH) Project.

Melissa M. Terras <m_dot_terras_at_ucl_dot_ac_dot_uk >, Department of Information Studies, University College London

Abstract

e-Science and high performance computing (HPC) have the potential to allow large datasets to be searched and analysed quickly, efficiently, and in complex and novel ways. Little application has been made of the processing power of grid technologies to humanities data, due to lack of available large-scale datasets, and little understanding of or access to e-Science technologies. The Researching e-Science Analysis of Census Holdings (ReACH) scoping study, an AHRC-funded e-science workshop series, was established to investigate the potential application of grid computing to a large dataset of interest to historians, humanists, digital consumers, and the general public: historical census records. Consisting of three one-day workshops held at UCL in Summer 2006, the workshop series brought together expertise across different domains to ascertain how useful, possible, or feasible it would be to analyse datasets from Ancestry and The National Archives using the HPC facilities available at UCL. This article details the academic, technical, managerial, and legal issues highlighted in the project when attempting to apply HPC to historical data sets. Additionally, generic issues facing humanities researchers attempting to utilise HPC technologies in their research are presented.

Introduction

Although HPC, pooled computational resources, shared large scale datasets, and associated "e-Science" [1] technologies have large potential for analysing and sharing complex data sets, little application has been made of the processing power of the grid or HPC technologies to furthering research on humanities data. This could be due to the lack of large-scale datasets requiring such technologies for analysis, and little understanding of or access to these e-Science technologies. The ReACH workshop series was established to investigate the potential application of e-Science and HPC technologies to one of the largest humanities and social science datasets in existence: historical census records. 1

Public interest in historical census data is phenomenal, as the overwhelming response to mounting the 1901 census online at The National Archives demonstrates [Inman 2002]. Yet the data is also much used for research by historians and social scientists (see [Higgs 2005] for an introduction). There are many versions of historical census datasets available, covering a variety of aspect of the census, and digitised census records are one of the largest digital datasets available in arts and humanities research. In the Arts and Humanities Data Service repository collection alone there are currently 155 datasets pertaining to historical census data (from the UK and abroad) created for research purposes [AHDS 2006]. Commercial firms dealing (or having dealt) in genealogy information (such as Ancentry, Genes Re-united, QinetiQ, British Origins, The Genealogist, and 1837Online) have digitised vast swathes of historical census material (although to varying degrees of completeness and accuracy). There is much interest from the historical community in using this emerging data for research, and developing tools and computational architectures which can aid historians in analysing this complex data (see [Crocket et. al. 2006]) for an advanced proposal regarding the creation of a longitudinal database of English individuals and households from 1851 to 1901 (see also the work of the North Atlantic Population Project). However, there have been few opportunities for the application of HPC to utilise large scale 2

processing power in the analysis of historical census material, especially analysing data across the spectrum of census years available in the UK (7 different censuses taken at 10 year intervals from 1841-1901).

3 The aim of the ReACH series was to bring together disparate expertise in Computer Science, Archives, Genealogy, History, and Humanities Computing, to discuss how e-Science techniques could be applied to be of use to the historical research community. The project partners each brought various expertise and input to the project. UCL School of Library, Archives and Information Studies, hosted the workshop series, having expertise in digital humanities and advanced computational techniques, as well as digital records management. The National Archives, who select, preserve and provide access to, and advice on, historical records, e.g. the censuses of England and Wales 1841-1901 (and also the Isle of Man, Channel Islands and Royal Navy censuses), were involved to provide access to and expertise regarding census material. Ancestry.co.uk, who own a massive dataset of census holdings worldwide, and who have digitized the censuses of England and Wales under license from The National Archives, were involved to provide access to digitised census material: the input of Ancestry was central to this research to gain access to the complete range of UK census years in digital format. Finally, UCL Research Computing, the UK's Centre for Excellence in networked computing, who have extensive HPC facilities available for use in research, provided much guidance and expertise regarding e-Science technologies to the project.

4 The ReACH project aimed to investigate the reuse of pre-digitised census data: presuming there was not funding available to be in the business of digitisation of other record data for any pilot project. (Additionally, the Library, Archive, and Arts and Humanities communities have been merrily digitizing resources in earnest for over twenty years: it was hoped that by analysing one of the largest available digitized datasets with HPC that the appropriation of e-Science technologies for humanities research could be demonstrated.) The project also wished to investigate the use of commercial datasets (as many of the large census data sets are owned by commercial firms: in this case, Ancestry), and the licensing and managerial issues this would raise. The project also wanted to establish how feasible, and indeed useful, undertaking such an analysis of historical census data would be.

5 The results of the well-attended workshop series were a sketch for a potential project, and also recommendations regarding the implementation of e-Science (HPC) technologies in historical research. However, at the time, it was not thought possible to pursue the potential project primarily due to the quality and scope of available historical data. This paper describes the methodology of the workshops, reports on suggestions made during the series, sketches out a future project regarding how historical census material can be analysed utilising HPC, and extrapolates recommendations that can be applied in general to the use of e-Science in the arts and humanities research sectors.

Methodology

Workshops

6 The ReACH project was based around a series of workshops which aimed to bring together cross-disciplinary expertise from industry, government bodies, and academia. All workshops were held at UCL in summer 2006. The workshops were split into three topics.

7 The All Hands Workshop aimed to ascertain how feasible, and indeed, useful utilizing e-Science technologies to analyse historical census data would be. Undertaking e-Science analysis of historical census records may be technically possible – but will it be useful to academic researchers? The workshop brought together a wide range of interdisciplinary expertise to ascertain the academic community’s view of the benefit and concerns in undertaking a full-scale research project utilizing available historical census data and the Research Computing facilities at UCL. Through various presentations and discussions, this workshop explained the technological issues and explored the historical techniques which may be useful for undertaking research of historical census material in this manner.

8 The Technical workshop built on conclusions from the All Hands Meeting. Participants were a smaller group of those from interested parties, meeting in order to ascertain the technical issues regarding mounting Ancestry and TNA’s historical census data on the UCL Research Computing facilities. This workshop meeting aimed to ascertain how the data will be delivered to UCL, the size of the data, the structure of the data, the function of searches to be undertaken

on the UCL Research Computing facilities, the duration of the project, the number and type of employees required, the equipment required to purchase, the equipment required to access to existing kit, the software required, software development issues, and other issues such as data security and management.

The Managerial Workshop was the final workshop to be undertaken as part of this research series. The aim of this workshop was to ascertain the managerial and legal issues which would need to be resolved in order to undertake a research project using Ancestry's data, in conjunction with The National Archives, and UCL. Issues which were discussed included; licensing requirements from Ancestry, security of data, ownership of research outcomes, management structure, financial structure, paths to dissemination and publicity, and other topics suggested by participants. 9

Follow Up to Workshops

Following the workshops, points of interest were pursued. These included checking reference material to understand prior research which had been brought to the PI's attention, making further links with other projects (such as the Centre for Local History Studies at Kingston University London which is constructing a comprehensive database detailing major aspects of Kingston's economic and social evolution during the second half of the nineteenth century) and the holders of other large scale data sets (such as the Free BMD Register which aims to transcribe the Civil Registration index of births, marriages and deaths for England and Wales). Individuals were also consulted from a diverse range of sources, including the Arts and Humanities Research Council's lawyers (who provided legal advice regarding the creation of new datasets through combining existing sources), the business development office at UCL, UCL's Centre for Health Informatics and Multiprofessional Education (who provided expertise on data security and management), and researchers in Physics working on the AstroGrid project (who were interested in seeing how results of a potential project could be useful for research involving scientific data). 10

Findings

Findings from the workshops are presented here, utilising the framework in which the workshops were presented, breaking the project into academic benefits, technical infrastructure needed, and management and legal issues which arose from the discussions. The following section, future work, details how the project could proceed in developing a pilot e-Science project in this area. 11

Benefits for Historians – Would this be useful?

There is significant interest in how HPC can aid historians in analysing, matching, and processing historical census data. Computational methods have been extensively used to clean, manage, manipulate and match census record holdings for decades (see [History and Computing 1992], [History and Computing 1994], [History and Computing 2006], [PRDH 2000], [Dillon and Thorvaldsen 2001], [Schürer & Woollard 2002] for just a small indication of the breadth of research available) but most processes are still dependent on human input on some part of the processing chain. The use of computational techniques also has been hampered by lack of processing power, lack of availability of data, and problems in quality of data. If there was unlimited processing power, which could be used to search and manipulate all of the UK historical census data in automated processes, should it be available in digital form, what could it do which would aid in the research of historians? The "wish list" for tools and processes that would aid historians and genealogists was extensive and varied. Some suggestions were more likely to be computationally implementable than others, but all are included here, disregarding computational complexity or reliance on available data. 12

The most popular request was the generation of automatic matches of records throughout the census years available, creating what is known as a "longitudinal database" of individuals across the census. This would require the investigation of tools, techniques, and algorithms, and modelling of procedures undertaken by historians when they carry out this task manually at present. It would result in a dataset which can be used historians to track individuals, families and population change across time, and inform other projects interested in building such datasets. 13

An additional aid to historians would be the generation of rich variant lists for users. The use of variants is important in 14

dealing with the problematic nature of census data, which can often have errors due to its nature of collection (see Findings, below). By building up lists of common variants present in the UK census data, this would help to normalise the lookup process for historians, and provide probabilistic information which could be used in any computer architecture created to match records. Lists of variants fall into a variety of categories: typographic (provo versus probo), phonetic (Cathy versus Kathy), cultural (the use of Jack for those officially named John), temporal (1880 written down when actually they meant 1881) and spatial (Boston, when Cambridge was the official answer). Using computers to automatically generate rich variant lists would be a relatively simple task, and of great use to historical researchers.

15 Computational tools could be used to check and cleanse census data. The 5% sample of 1881 census data digitised and developed by Kevin Schürer and Matthew Woollard [Schürer & Woollard 2002] required a program of "enrichment" to reformat input data, perform a number of constituency checks, and add a number of enriched variables, mainly relating to household structure [Schürer & Woollard 2002, 16]. Manually checking a dataset of this size (around one million records) was not feasible, and "automatic validation and enrichment of the data is intellectually more rigorous than manual intervention" (ibid) whilst ensuring that the data is consistent across the dataset. The processing power necessary for running such algorithms across the whole of the UK historical census data and across each UK census is large and would require that afforded by e-Science technologies: 29 million records (or so) per census, and 7 census years (1841-1901). (See [Schürer & Woollard 2002, Appendix C] for a detailed discussion of the procedures carried out in their study.)

16 Calculating and identifying individuals who have been missed in various censuses is also possible. These may be individuals who were not "at home" on the night the census was taken, or those who were homeless, in mental institutions, etc. Identifying and calculating individuals who are missing from the census is a concern for modern day statisticians. (In the 2001 census, for example, it was estimated that a significant number (600,000) of young men, in particular, had disappeared from the statistics, and were unaccounted for [BBC 2004].) This could be revealed through longitudinal studies – and also provide further information about the quality of the census data itself.

17 Missing data in the digital records can be reconstituted through contextual information: for example, street numbers are missing in the Ancestry dataset, but could this be inferred from the surrounding data, allowing us to construct richer datasets looking at surrounding records? Can the number of rooms in dwellings be calculated? Reconstituted and enriched datasets can be useful to historians, provided that original transcripts are maintained and data integrity preserved for quality control, as in the enriched dataset in [Schürer & Woollard 2002]).

18 If digital data is held for all censuses, it can be used to generate simple statistics regarding the number of records for each parish. These results were previously published just after each census was collected in population reports (which are now being digitised themselves by the Online Historical Population Reports Project) and contain detailed analysis of the census results without naming individuals: for example, the reports give overviews of the size of parishes (geographically), the number of households, the number of male and female persons, numbers of male and female persons under 20 and over 20, etc. These statistics were calculated manually from the enumerator returns. It would be possible to check the accuracy of these by automatically counting the same fields in the digital records for each census. This, of course, could also be used to check the accuracy of the digital records: any discrepancies between the two would have to be investigated.

19 A popular, yet computationally difficult, suggestion of facilities that would help researchers was the development of OCR techniques which can be used effectively on copperplate handwriting, in order to be able to digitise missing fields quickly and efficiently. (For example, the occupation field was missed from the Ancestry digitisation procedures to cut digitisation costs, but occupation data is one which is most often used by historians). Research into automatic optical character recognition of handwriting, although extensive, has yet to generate techniques with a high enough success rate to allow this to be a feasible project at this time (see [Impedovo 1993]) for an accessible overview of techniques and approaches commonly used).

20 There was interest in using computers to map census data onto geographical information. Firstly, a popular suggestion was the name mapping of geography to names. There has been some success with this – a UCL project based in the

Centre for Advanced Spatial Analysis has been working on a Surname Profiler which investigates the distribution of surnames in the UK in both historic (1881) and contemporary (1998) census datasets. (A conference regarding the benefits this has for research was held at UCL 28th- 31st April 2004. See [Lloyd et. al. 2004] for an overview and collection of papers presented.) Extrapolating the research across all the census years will require much processing power, firstly, to enable the cleansing and formatting of the data, secondly, to allow generation of results, and finally, to increase the sophistication of visualisation techniques to show the changing of distribution of surnames through time.

There was also great interest in assigning grid references to historical data. The boundaries of districts, and indeed, names and areas of census parishes differ greatly from census to census (see [Mills et. al. 1989] for an overview of related research). There is currently no way of automatically relating places which appear within one parish in one census and another in the next. This makes automated linkage of records difficult. Investigating how geo-spatial references can be applied to individual areas across the spectrum of census data will allow new datasets to be formed which can aid historians in tracing how settlements have changed, irrespective of the changes in legislative boundaries. 21

Related to this was the request for the addition of current geographical data to the census. It is a common request at the National Archives for people to be able to search historical census data on current postcodes. Although this will be a complex and difficult endeavour (many street layouts have changed, postal districts and boundaries change, and the attempt will require a thorough understanding of urban geography from 1841 onwards, which may be impossible to model computationally) this tool would be welcomed by, in particular, family historians and genealogists. 22

Visualisation techniques could be employed to investigate how the data was collected, the distribution of different fields across the geography of the UK, and the way that the distribution of data changes from census to census. If geo-spatial data were to be generated, or become available, it could be manipulated through GIS, increasing the means to interrogate and conduct new research with the data. (Visualisation of scientific data has been a focus of the use of e-Science technologies within the mathematical and physical sciences, see [Brodlie et. al. 2004] and [Riedel et. al. 2007], although again, use of these advanced computational techniques in the humanities is nascent.) 23

A practical suggestion was for the generation of tools which can be used for social computing – looking at family histories as opposed to individual histories, to investigate family roles and structures across the different census years, which would be a useful practical addition for those carrying out genealogical research. 24

Finally, separate from the analysis of the data itself would be the facilities to analyse how people were actually using the data: it is known to be popular, but not much more is known about how people search, analyse, and link census material. Log analysis of usage statistics from those accessing historical census data online could be undertaken to provide quantitative evidence regarding use, which would be useful to understand the nature of genealogical research, and also the procedures used to match records. (See [Nicholas et al 2006] and [Huntington et. al. 2007] regarding how these techniques have been employed to understand digital user behaviour of other online resources, and [Warwick et al. 2008] for this technique applied particularly to those in the arts and humanities.) The popularity of the census material would mean that logs generated from the searching of these records would be large, and require the facilities afforded by e-Science technologies for efficient analysis. 25

But where is the “e-Science” in all this? Most of these projects would require large processing power, to begin to sort through the large dataset. Mike Mansfield, on 14th June, informed us Ancestry has approximately 600 Tera-Bytes of census data holdings, including image files [Mansfield 2006]. The English, Wales and Channel Islands textual data for 1841-1901 is a mere 20 Giga-Bytes in comparison: with over 200 million individual records to perform some kind of task on. Manipulating that volume of records as one dataset requires processing power not readily available in a desktop machine. The more complex the task, the bigger the data storage (both for temporary data manipulation and for storing results) required. Although this dataset is a lot smaller than most of the datasets scientists at UCL are using in their e-science projects (see http://www.ucl.ac.uk/research-computing/ for an overview), making use of the HPC facilities at UCL would allow this data to be interrogated in a reasonable and realistic timeframe. 26

However, whether using HPC to manipulate data is actually “e-Science” is open to question. The AHRC’s definition of e-science varies somewhat, but is stated on their webpage as 27

> a specific set of advanced technologies for Internet resource-sharing and collaboration: so-called grid technologies, and technologies integrated with them, for instance for authentication, data-mining and visualization. [AHRC 2006]

and in a presentation introducing the e-Science main call for funding more succinctly as

> the development of *advanced* technologies for research collaboration and resource sharing across the Internet.
>
> - Grid technologies, and technologies integrated with them (*service* grid)
> - Not e-Research
>
> [Robey 2006]

This raises larger questions about what e-Science actually is, and whether the development of new advanced high performance techniques would fit under this rubric. (Although research should be problem and solution led, rather than definition led, funding is required to carry out research of this type.)

It is doubtful whether a project regarding processing of census data would either need to use (or be wise to use) computational grid technologies to undertake its processing (see Technical Implementation below). Processing would be carried out by a high performance machine, not dispersed across the computational grid (why make the project more complex than it needs to be?) There are additional security problems in sharing processing and datasets across the computational grid, or making them available via the National Grid Service, or even the Internet. When dealing with commercially sensitive datasets such as the census data from Ancestry, the value of that data should be respected (and the potential consequences of leaking this data to the world realised): therefore, constraining the processing of the data to one individual system is advisable, rather than copying and distributing it over a network, which provides a higher chance for interception and malicious (or other) copying and unlawful dissemination. Thus, any project would not be "e-Science" in this regard: as the data would not be distributed, or made more available than it currently is to those not part of the project. 28

Finally, the question of the ownership of any newly created datasets from the programme is tricky, as is the extent to which the commercial data is part of these datasets, or compromised by sharing the datasets (see Managerial Issues below). Therefore, distribution of the *results* of the project may not be possible via Internet or Grid. 29

The potential for (the AHRC's definition of) e-Science when dealing with commercially sensitive data is therefore much reduced. In the future, as more datasets are being created in the public domain, this will become less of a problem as researchers should not have to rely on commercially provided data. 30

A further important topic that was discussed in the All Hands Meeting was the quality and integrity of historical census data. This is reported on below in Future Research, and issues regarding data security and procedures are covered in Managerial Issues. 31

Technical Implementation – Would this be feasible?

In many respects, technical implementation of a project which would input Ancestry's datasets, perform data manipulation, and output data, is much less of a problem than identifying the research question, due to the excellent research computing facilities and support available at UCL. Discussion regarding the range of expertise, services and facilities on offer is available at http://www.ucl.ac.uk/research-computing/information/services/, but can be summarised as AccessGrid facilities for virtual collaboration, Central Computing Cluster (C^3) for advanced batch style computing, e-Science Certification for use of national grid resources, Condor high-throughput commodity computing pool, Prism high-performance visualisation resource, The Sun Cluster "Keter" for serial and parallel computing, and the Altix for High-Performance Computing [UCL Research Computing 2006b]. 32

For the security reasons outlined above and in Managerial Issues below, any project would have to use a standalone 33

machine rather than distribute data via a network (such as the Condor computing pool) for processing via a grid or grids. After consultation with UCL Research Computing regarding memory requirements, scalability and Input/Output (I/O) profile, it was determined that the SGI Alrtix facility at UCL (one of two facilities for parallel computing, the other being the Keter cluster) would be the most suitable choice, with 56 processors (Itanium2 1.3Ghz/3 MB cache processors) and 112GB shared memory offering speeds of approximately 135GFlops [UCL Research Computing 2006a]. Although various end-user packages are already installed on this system, the project would require development of its own software. The Altix facility has Intel C/C++ Compilers versions 7.1, 8.1 and 9.0 installed, and so would support C++ 20 based programs: a standard in the development of software tools.[2] C++ routines could be developed in a normal offline development environment, and sent to the Altix as a series of parallel jobs which would process the data. Obviously, this would require employing a programmer with prerequisite experience and abilities who could write C++ code for this project. The difficulty lies in *what* to program.

Because of security issues, data would be received from Ancestry on encrypted physical media rather than being transferred via Internet Technologies such as FTP. This would then be uploaded to the Altix when needed, whilst ensuring robust security measures were kept in place. Research Computing at UCL has much experience regarding data integrity and security with its many projects which carry out medical research such as those based at the Centre for Health Informatics and Multiprofessional Education (CHIME). Other projects using UCL's research computing facilities which require close management of ethical and security include the Co-operative Clinical e-Science Framework (CLEF), which looks at, amongst other things, security and privacy of clinical data. Recommendations regarding security procedures are made in the following section. 34

Likewise, temporary data storage facilities to allow processing would have to be secure, as would the storage of the results of the project. In many ways this is a simple I/O processing task: it is just the volume of the data, and the potential complexity of any developed algorithms which require high processing computing. There are no technical barriers to proceeding with this manner, and the facilities at UCL are even available free of charge for research to all UCL departments. 35

Managerial Issues – Would This Be Achievable?

Managerial issues of a potential, distributed project, fall into a variety of topics. Firstly, the managerial structure of the project. Secondly, management of security of data whilst the project is underway. Finally, ownership of results (whether datasets or algorithms) is of utmost concern in a project such as this which incorporates commercial partners: no-one wants to be exploited. 36

Management structures in projects such as these are fairly standard. A Principal Investigator from the Research institution would be responsible for the project overall, maintaining regular contact with the partners, having regular meetings, and reporting at regular intervals. An interdisciplinary steering committee is also advisable, to ensure all aspects of computation and historical interest would be represented. Regular meetings and updates are essential, as is the maintenance of documentation, and information provided publicly such as through a website. On an individual level, Research Assistants (particularly the programmer) should keep lab books regarding progress. All code should be commented, and documented. Backup procedures should be undertaken regularly. 37

Security issues regarding dealing with commercially sensitive data need to be resolved before delivery of data is made. Consultation with data management expertise in CHIME resulted in the recommendation of ISO/IEC 7799:2005, a comprehensive set of controls comprising best practices in information security which is an internationally recognized generic information security standard [ISO 2005]. Far from being an impenetrable managerial report, the standard sets out guidelines and principles regarding initiating, implementing, maintaining and improving information security and is commonly used when dealing with ethically or commercially sensitive data. Most importantly, the standard aids in setting up "effective security management practices, and to help build confidence in inter-organizational activities" [ISO 2005]. Establishing policies and procedures which can be agreed with partners in advance of providing data will foster trust – and aim to protect the partner in the project to whom the data is being supplied. Other measures which should be undertaken is maintaining a register of assets, obtaining secure off-site backup, undertaking thorough risk analysis, and 38

maintaining a “no surprises” approach to data flow to ensure good practice. Useful relevant literature regarding risk analysis, data management and systems security include [Adams 1995], [Anderson 2001], [Stallings 2005], [Stallings 2008].

39 Legal agreements should also be undertaken about the fair use and application of data for the duration of a project, and what happens to the data after the project ends. This will require legal assistance from institutional lawyers (who often provide the service free to the project on behalf of the institution: if this is the case the project need not include legal costs in its budget). It should not be underestimated how long it would take to draw up these agreements.

40 There are also considerations that need to be made regarding what happens to data resulting from the input of many project partners at the end of a project. Issues of longevity, preservation, and sustainability of research results are important, especially since the announcement that the AHRC will no longer fund the Arts and Humanities Data Service, where projects would have previously deposited their data to ensure long term access. More seriously, though, is the issue of who would own the resulting new data sets created as part of a project, or intellectual property rights on algorithms developed. Advice was taken from the AHRC Research Centre for Studies in Intellectual Property and Technology Law at the University of Edinburgh on this matter. There is currently much discussion in the legal field on the use of data in the research sector, and how the IP rules can best be used to support the aims of the teaching and learning community (see [Davies & Withers 2006] for an overview). If a new database of results was created in a potential project, the underlying rights would seem to fall under the protection accorded by the Database Directive [European Parliament 1996] (although further legal advice should be taken on this as copyright may also be an issue). Broadly speaking, each of the institutions who contributed data may have the database right in the contents of their databases. Where a substantial part of the source database is used, then permission would be needed to extract and re-utilise the contents. A “new” database right would result if these were combined for other purposes: the right would reside in the person or organisation who made an investment (whether it be financial, or time and effort) in compiling a new database. Much might depend on who was using or going to use the resultant product (for example, use may be limited to research and education). It is important that these questions are resolved at the outset of a project, to enable researchers to use and publish results, protect the commercial rights of the company, and also protect the intellectual investment of the researchers, especially regarding any outcomes which may be suitable for knowledge transfer or technological spin-off.

41 In a case such as the proposed project with historical census material, suitable agreements and licenses would have to be drawn up between all parties prior to the research commencing. In response to gaining access to Ancestry’s datasets, for example, UCL could grant Ancestry a time limited license for application of research results with the genealogical market. The researchers should be careful not to sign away rights to research outcomes.

42 For a grant application, it is important to establish managerial principles which will be resolved prior to the grant commencing, and to make sure that the institution has infrastructure to support these legal issues. The technology transfer office, or business office, at most universities will have expertise in this (usually in the scientific domain, but these procedures will also be applicable in the arts and humanities). UCL Business PLC was contacted, and advised the standard procedures for setting up a project was to establish the following: that the Researcher and UCL will retain the right to publish, that UCL Business PLC, and the Contract Research Office, will arrange IPR agreements and commercial exploitation, that the foreground IP of the project will remain the property of UCL, that commercial background IPR (data, etc.) will be licensed accordingly, that UCL Business and the related infrastructure can assist in all of this, and finally, that standard Data Protection procedures should also be applied. It was also stressed that adequate time should be given to resolve licensing and technical matters prior to a project commencing.

43 The barrier to setting up a project regarding processing of historical census data is not managerial: although it would take time on the part of the partner institutions to come to legal agreements regarding access to and sharing of the data. Many institutions have procedures in place to deal with such projects. Negotiating such licenses may take up a large portion of time at the outset of a project, however, and academic researchers should be prepared to come to grips with the intricacies of digital copyright and database law.

Future Research?

Following consultation with historians, it was obvious that the most popular, useful, and popular, project to pursue from this research would be one that looked into the techniques and procedures used to create longitudinal databases – tracking and tracing individuals and families across different census years, and enabling historians to look at the "life histories" of individuals, families, and properties. By investigating these procedures, using the available datasets, and implementing techniques which could use the processing power of UCL's HPC facilities (meaning that computational time would not be of concern to the project) it may be possible to undertake a comprehensive review of previous techniques used to carry out record linkage across the census, develop and implement new, robust procedures and techniques to undertake automated record matching using HPC across fuzzy datasets, and develop tools for historians undertaking the construction of longitudinal datasets, to aid them in checking and investigating possible linkages across datasets. 44

The knowledge transfer opportunities from developing robust and benchmarkable techniques would be large: consultation with Physicists working on the AstroGrid , for example, revealed that they are facing the same problem: being able to track and trace individual entities across fuzzy and incomplete datasets. Datasets from local, and central, governments have the same problems, as do matching individuals across credit records in the financial sector. Moreover, the development of tested techniques would further the aims of historians in being able to create longitudinal datasets, and would be of great interest to genealogists, and companies operating in the genealogy sector. The results from such a project would sit alongside, and feed into, Crocket, Jones, and Schürer's proposed Victorian Panel Study Project (2006). 45

However, the problem of automatically matching individuals across census years is not trivial. Firstly, the nature of census data is that quality will always be of concern to the historian, and matching records across years therefore deals with great levels of uncertainty. There has been much research into the inherent qualities of census data (for example, see [Holmes 2006], for an investigation into common problems in Ancestry's datasets. Other relevant research includes [Tillot 1972], [Perkyns 1991], [Perkyns 1993] and [Woollard 1997]). Errors in the data can be introduced at every level: from those supplying the data (who may not have known, for example, how to spell their name or their precise birth date), from those writing down and transcribing those answers into the enumerator returns, and from those entering the data into the digital version of the records. This is something that has to be accepted when dealing with census data. Although computational methods can be use to check data quality and normalise some discrepancies (see [Schürer & Woollard 2002]), census data will remain "fuzzy", and often incomplete. This makes computational matching of data difficult. 46

Added to this is the problem that the digital datasets themselves may not have certain fields digitised (depending on the digitiser, often important fields of data are missed to cut digitisation costs. The Ancestry datasets, for example, do not have occupation digitised, which can often be used as an indicator of identity). Without the full data available across the UK, it is difficult to develop algorithms or procedures which can undertake record linkage across the data. 47

Ten years elapse from census to census – people can move, marry, remarry, be born, die, or change name. Techniques used to match individuals from census to census usually depend upon having other data available to "triangulate" individuals – for example civil registers such as births, deaths and marriages, or parish burial records. Often projects have to digitise the material themselves, as it is not often in the public domain (the FreeBMD project aims to transcribe the Civil Registration index of births, marriages and deaths for England and Wales, and to provide free Internet access to the transcribed records – although this is very much work in progress, dependent on volunteer labour). An example of a project utilising these different information sources to undertake longitudinal analysis of historical census data is the Cambridge Group for the History of Population and Social Structure, which has created 48

> Four parallel longitudinal data sets...by linking individuals in the decennial censuses of 1861-1901 with the births, deaths and marriages from civil registers for the lowland town of Kilmarnock, the Hebridean Island of Skye, and the rural parishes of Torthorwald and Rothiemay, places with contrasting economic and social structures and physical environments. [CamPop 2006, 26]

This work was dependent on them being granted special permission by the General Register Office, Edinburgh, for access to the civil registers of births, marriages and deaths, and the creation of database of this material by a project worker.

49 The Kingston Local History group is also interested in linking records across the different census years, and is constructing

> a comprehensive database detailing major aspects of Kingston's economic and social evolution during the second half of the nineteenth century. The core of the database is the complete census enumerators' returns for each census year 1851-1891 (145,000 records). [Kingston Local History Project 2006]

The project is dependent on four datasets, which they have either constructed or cleaned up: The Census Enumerators Books for Kingston Upon Thames Census Area - 1851 to 1891; the Bonner Hill Cemetery Burial Registers - 1855 to 1911; the Kingston Parish Burial Registers - 1850 to 1901; and the Kingston Parish Marriage Registers - 1850 to 1901. This obviously required much investment to allow the project to undertake research (see [Tilley 2003a] for further details of the project). Peter Tilley has also developed computational tools to allow record linkage to be undertaken, including some formalisation of the procedures historians use to undertake record linkage (see [Tilley 2003b] for an overview of these tools and research generated from them). The tools are not wholly automated, though – which would be necessary to generate a full scale longitudinal survey of the English historical census.

50 Even with these difficulties, there is much interest in the possibilities of Automated Record Linkage techniques for linkage of census data (see [History and Computing 1992], [History and Computing 1994] for an overview of research) in particular for the automated tracking and tracing of individuals across the different census years to produce what is known as a Longitudinal data set (see [History and Computing 2006], and the Victorian Panel Study [Crocket et. al. 2006]). Projects have difficulties in two areas: projects are still dependent on human interaction in the data linking routines, meaning that routines are not wholly automated, and projects are dependent on the creation of more datasets (Campop, Kingston, and VPS.)

51 Only when in depth datasets from across the UK are available will it be possible to carry out a full scale longitudinal survey: although there has been much financial, industrial and academic investment in the creation of digital records from historical datasets, there is not the quantity nor quality of information currently available to allow useful and usable results to be generated, checked, and assessed from undertaking automatic record linkage in this area.

52 However, one of the aims of the ReACH project was to investigate *re-use* of digital data. The Kingston project has digitised material, and constructed a longitudinal dataset through the input of researchers regarding the area of Kingston Upon Thames in the Victorian era. A potential project lies in taking this dataset, and its constituent forms, and *trying to recreate this data set computationally*, thus being able to test and benchmark procedures against a “quality controlled” dataset. If computational algorithms can be developed which are as effective as a human researcher in creating linkage across this relatively small dataset, then perhaps these could be scaled to cover the whole of the English data when it becomes available. Moreover, certain subsets of the data prepared by the Kingston team could be replaced at certain points in the project with other datasets – such as the Ancestry data from the same area – to investigate whether it would ever be possible to scale the project up using these pre-digitised datasets which had not been digitised for the purpose of record linkage.

53 How would such a project proceed? A process of knowledge acquisition (conventionally defined as the gathering of information from any source) and Knowledge Elicitation[3] (the subtask of gathering knowledge from a domain expert, see [McGraw & Harbison-Briggs 1989], [McGraw & Westphal 1990], [Shadbolt & Burton 1990]) would have to be undertaken, regarding both how experts carry out record linkage manually, and a literature survey on previous research undertaking automated record linkage (not only restricted to that regarding historical census data). The available literature on automatic record linkage is large and expansive 28, and the techniques already attempted and employed not trivial: it will be difficult to contribute anything meaningful to this vast body of research 29. Given the amount of prior research in the area, the chances of developing novel architectures and algorithms which can carry out record linkage

scaleable over the available census data is slim. Access would have to be gained to a range of data (in this case Kingston datasets, but also those from Ancestry, Free BMD, and any other of the relevant census datasets which are available) and licenses for use and contracts negotiated for each. A secure system would have to be set up to receive and store data. Programming the potential techniques would then begin, constructing a system which would which mount data, apply techniques, and output results. Tools and techniques for monitoring and benchmarking the quality of these results against the test datasets would have to be established. Finally, results would have to be published and disseminated.

54 It is obvious from this outline that this would project will take some time and manpower to carry out. It is estimated that a three to four year project featuring one historian/knowledge engineer and one computer scientist, as well as input from the Principal Investigator, and involving consultation with many historians, should be able to undertake this work. Initial costings suggest this would be very expensive, however. The project is also very “blue-sky”. It may not be possible to automate the record linkage routines adequately, nor develop any automated record linkage techniques which are more effective than those which currently exist, or scale the results up at the moment due to the lack of existing datasets of quality, making this a potentially lengthy and costly exploration with a high risk factor.

55 Unfortunately, should a record linkage project be carried out on the Kingston Upon Thames area data, developing routines which could be checked against the database which has already been constructed and checked by researchers is, at current time of writing, not available to allow results to be scaled up to the rest of the country. Births, marriages and death indexes are not fully available or digitised, and due to the economic climate of Kingston in the Victorian era, it can be argued that results from such a stable, middle-class environment would not be applicable to other, very different parts of England. Although the potential project is interesting, and could develop new algorithms for automated record linkage which could be checked and benchmarked against a human constructed linked database of quality, it was decided that at the current time, with the available data, that the low chance of obtaining positive research outcomes from the project would not balance the financial and intellectual investment required to undertake the research.

Findings: e-Science and the Humanities

56 Undertaking the ReACH series has resulted in various findings and recommendations which can be useful to other projects in the research field, but also useful to those considering using (or even funding!) e-Science or HPC technologies for humanities research.

For the Historian

57 There were various points of note for historians. Firstly, although there has been much financial, industrial and academic investment in the creation of digital records from historical census data, there is not the quantity nor quality of information currently available to allow useful and usable results to be generated, checked, and assessed from undertaking automatic record linkage across the full range of census years. If the project above were carried out on a subset of the census data, results would not yet be scaleable across England due to lack of data currently available. This will change as more data is digitised (and becomes available to the general research and genealogical community through publicly available websites operating under appropriate usage licenses). Secondly, the potential for high performance processing of large scale census data is large, and may result in useful datasets (for both historian and genealogist) when adequate census data becomes available. This should be revisited in the future. Access to computational facilities or expertise or managerial issues were not the limiting factors here (at least at UCL, although it is understood that other institutions may not have such easy access to such infrastructure).

For the Arts and Humanities Researcher

58 Generic issues raised which may be of interest to researchers in e-Science and the Arts and Humanities include the fact that the HPC and e-Science communities are very welcoming to researchers in the arts and humanities who wish to utilise and engage with their technologies. There is also potential for research in the arts and humanities informing research in the sciences in this area, particularly in areas such as records management, information retrieval, and dealing with complex and fuzzy datasets.

The problems facing e-Science research in the arts and humanities are predominantly not technical. Although there is still fear in using HPC in the arts and humanities, dealing with the processing of (predominantly) textual data is not nearly as complex as the types of e-Science techniques (such as visualisation) used by scientific researchers. However, the nature of humanities data (being fuzzy, small scale, heterogeneous, of varying quality, and transcribed by human researchers) as opposed to scientific datasets (large scale, homogenous, numeric, and generated or collected/sampled automatically), means that novel computational techniques need to be developed to analyse and process humanities data for large scale projects, and often large enough data sets of high enough quality which warrant the use of these technologies are not available. 59

Using the processing power of computational grids may be unnecessary for humanities projects if data sets are small, and projects have access to stand-alone machines which are powerful enough to undertake the task themselves. Processing data via computational grids can be a security risk: the more dispersed the data, the more points of interception there are to the dataset. Researchers should choose the technologies they use to carry out processing according to their need, but often running queries on a stand-alone high performance machine requires less managing at present than using processing power dispersed over a local, national, or international grid. Additionally, the challenging nature of humanities research questions mean that they are often not predisposed to batch processing and running as repetitive jobs. 60

Finding arts and humanities data which is of a large enough size to warrant grid or high performance processing whilst being of high enough quality can be a problem for a researcher wishing to HPC in the arts and humanities. This may just have to be accepted, and the fuzzy and difficult data generated regarding arts and humanities data explored and understood to allow processing to happen. In this way, using e-Science to deal with difficult datasets could benefit computing science and internet technologies too. Perhaps this is the main thrust of where e-Science applications in the arts and humanities may have uses for others – and knowledge transfer opportunities. 61

Where commercial and sensitive data sets are involved in a research project, Intellectual Property Right issues and licensing agreements should be specified before projects commence. The importance of this issue cannot be stressed enough – especially when the project is wholly dependent on receiving access to datasets, or dealing with commercially valuable and sensitive data. Commercial companies are often keen to be involved in research if there are benefits to themselves: nevertheless, the IPR of academic institutions should be safeguarded. This can best be achieved through setting up specific licenses for the use of algorithms in the commercial world: again, this should be ascertained before the project commences. 62

Those in arts and humanities research may not be used to dealing with legal aspects of research. Most universities have legal frameworks in place to deal with such queries in the case of medical and biomedical research. These facilities are generally available free of charge to arts and humanities projects within their institutions, and so funding would not be compromised by having to include legal charges in funding bids. The time taken to negotiate licenses for data use should not be underestimated, however. Advice should also be taken from those involved in biomedical research: the similarities between projects in this area and the arts and humanities are significant when it comes to data management, IPR, copyright, and licensing issues. In particular, where sensitive data sets are used, the arts and humanities researcher should look towards medical sciences for their methodologies in data security and management, in particular utilising ISO 17799 to maintain data integrity and security. 63

For Funding Councils

Where e-Science arts and humanities projects involving large datasets are proposed, it is likely that the complexity of the project will require large scale funding. Yet many of these projects will be “blue-sky”, and may require a variety of employed expertise over a number of years to undertake the work, as well as requiring technical expertise and infrastructure. These projects will then be expensive: funding calls in e-Science for the Arts and Humanities should take this into account. 64

Additionally, e-Science projects in the arts and humanities may be high risk with less definable outcomes than similar projects in the sciences, due to the complexity and inherent qualities of arts- and humanities-based data. If funding 65

councils wish to foster success in this area, the risks of funding such projects should be acknowledged. The very attempt to develop *practical* projects which wish to apply e-Science technologies in the arts and humanities may result in cross fertilisation with the scientific disciplines.

Definitions of e-Science vary from council to council. HPC is as much a part of “e-Science” in the sciences as distributed computational methods, yet the definition of e-Science for the arts and humanities focuses on networked computational methods. The two should not be distinguished from each other. If there are to be different definitions of e-Science between the arts and science councils, the reasons for this should be researched and expressed to elucidate different funding council’s approaches to e-Science, and to further explore where e-Science technologies can be of use to arts and humanities research. 66

Conclusion

The ReACH workshop series has successfully brought together disparate expertise on history, records management, genealogy, computing science, information studies, and humanities computing, to ascertain how useful or feasible it would be to set up a pilot project utilising e-Science technologies to analyse historical census data. 67

There was much interest in the series, as the topic of how HPC facilities can be embraced by the arts and humanities audience is a pertinent one: funding for e-Science facilities is now becoming available for researchers in the arts and humanities, but how can these be appropriated by the domain? 68

An interesting aspect to the workshop series was defining the research question. Datasets were available, expertise was available, and unlimited processing power was available – but could these be harnessed to provide a useful and useable product for historians? The “wish list” from historical researchers is illuminating, indicating the potential for HPC in this area if and when comprehensive data sets of high enough quality become available, although they do demonstrate that novel, advanced computational approaches may have to be developed to deal with the real world complexity of humanities research questions and complex humanities datasets. 69

Aspects which may be peculiar to this project regarding collaborating with commercial partners indicate the managerial and legal similarities between research in the sciences and that in the arts and humanities. Researchers in the arts and humanities may find it useful to make contact with those in the sciences to ascertain which procedures are commonly undertaken in these areas. An interesting difference between the two, though, is the nature of humanities data, versus scientific data, which has been somewhat explored in this project. Whereas scientific data tends to be large scale, homogenous, numeric, and generated (or collected/sampled) automatically, humanities data has a tendency to be fuzzy, small scale, heterogeneous, of varying quality, and transcribed by human researchers, making humanities data difficult (and different) to deal with computationally. However, ascertaining how large scale processing of this type of data can be undertaken will be useful for computer science: if procedures for dealing effectively with difficult and fuzzy data can be resolved, these can be applied to a range of computational activity out-with the arts and humanities domain. Tackling e-Science projects in the arts and humanities may then inform developments in computer science for other applications. 70

Although the ReACH series came to the conclusion that the time was not right to carry this project forward into a full scale funding proposal and project, it is hoped that the findings of the workshop series will be of interest to others wishing to apply high performance processing to large scale humanities datasets. e-Science technologies still have the potential to enable large-scale datasets to be searched analysed, and shared quickly, efficiently, and in complex and novel ways: developing a practical project which explores humanities data in this manner should be rewarding for both humanist and scientist alike. 71

Acknowledgements

The ReACH project involved many individuals from a range of academic backgrounds, and the project would not have been a success without the input from project partners, those attending the workshops, those who provided advice and support when approached, and those on the steering committee. 72

Josh Hanna (Ancestry.com), Ruth Selman, and Dan Jones (both National Archives) all provided the project with their expertise. The project is particularly indebted to Jeremy Yates and Clare Gryce, both from Research Computing, UCL, for their continued input and support.

The speakers from the first workshop were Clare Gryce (Research Computing, University College London), Ruth Selman (Knowledge and Academic Services Department, The National Archives), Keith Cole (Census Data Unit, National Dataset Services Group, MIMAS, The University of Manchester), Ros Davies, Eilidh Garrett and Alice Reid (Cambridge Group for the History of Population and Social Structure) Mike Wolfgramm (Vice President of Development, MyFamily). The success of the workshop was dependent on their presentations, and follow up discussions, and the project appreciated their involvement.

Participants of the various workshops, who were responsible for lively discussion and intellectual input into the project, included Kevin Ashley (Head of Digital Archives, University of London Computer Centre), Tobias Blanke (Arts and Humanities e-Science Support Centre), Keith Cole (Director of the Census Data Unit, Deputy Director of National Dataset Services Group, MIMAS, The University of Manchester), Ros Davies (Cambridge Group for the History of Population and Social Structure), Eccy de Jonge (Research Administrator, UCL SLAIS), Matthew Dovey (Technical Manager, Oxford E-science Centre, University of Oxford), Eilidh Garrett (Cambridge Group for the History of Population and Social Structure), Clare Gryce (Manager UCL Research Computing, Department of Computer Science, UCL), Josh Hanna (Managing Director and Vice President, Ancestry Europe), Edward Higgs (Reader, Department of History, University of Essex), Richard Holmes (MA Research Student, UCL), Dan Jones (Licensing Manager, TNA), Andrew MacFarlane (Lecturer, Department of Information Science, City University), Duncan MacNiven (Registrar General for Scotland), Mike Mansfield (Director of Content Engineering and Search, MyFamily Inc), Pablo Mateos (Department of Geography / CASA, University College London), Gill Newton (Cambridge Group for the History of Population and Social Structure), David Nicholas (Chair of Library and Information Studies, UCL SLAIS), Chris Owens (Head of Access Development Services, The National Archives), Rob Procter (Research Director of the National Centre for e-Social Science), Alice Reid (Cambridge Group for the History of Population and Social Structure), Kevin Schürer (Director of the Economic and Social Data Service (ESDS) and the UK Data Archive (UKDA), Department of History, University of Essex), Ruth Selman (Knowledge and Information Manager, The National Archives), Leigh Shaw-Taylor (Cambridge Group for the History of Population and Social Structure), Edward Vanhoutte (Co-ordinator, Centre for Scholarly Editing and Document Studies (KANTL), Ghent), Claire Warwick (Lecturer in Electronic Communication and Publishing, UCL SLAIS), Jeremy Yates (UCL Research Computing, Lecturer in Physics and Astronomy, UCL), and Geoffrey Yeo (Lecturer in Archives and Records Management, UCL SLAIS).

Following the workshops, various individuals provided further advice. Anna Clark and David Ashby (both UCL Business) provided legal advice; Charlotte Waelde (AHRC Research Centre for Studies in Intellectual Property and Technology Law, School of Law, University of Edinburgh) also provided advice on legal matters. Nathan Lea (UCL Centre for Health Informatics & Multiprofessional Education) provided advice on data security and management.

Peter Tilley and Christopher French (both Centre for Local History Studies, Kingston University London) provided advice and were keen to collaborate on future research projects, offering access to the data which has emanated from their research projects. Ben Laurie (FreeBMD) also offered his support for the project, and was keen to collaborate further.

The steering committee comprised of Tobias Blanke (Arts and Humanities e-Science Support Centre), Alastair Dunning (Arts and Humanities Data Service), Lorna Hughes (AHRC Methods Network), Dolores Iorizzo (Centre for the History of Science, Technology and Medicine, Imperial College London), Martyn Jessop (Centre for Computing and the Humanities, King's College London), Dan Jones (The National Archives), David Nicholas (UCL SLAIS), Kevin Schürer (University of Essex), Ruth Selman (The National Archives), Matthew Woollard (History Data Service), and Geoffrey Yeo (UCL SLAIS).

The project would especially like to thank Tobias Blanke for his support and enthusiasm, Matthew Woollard for his sound advice, and Eccy de Jonge and Kerstin Michaels, both UCL SLAIS, who provided the project with excellent administrative support. Final thanks to Andrew Ostler for his support.

Notes

[1] "e-Science" is a term given to a variety of technologies covering high performance, large scale, and grid enabled computing, and the shared data and computational resources used in these technologies. See [Dunn & Blanke 2009] in this issue for an overview of the use of this term, and the US equivalent, *cyberinfrastructutre*.

[2] C++ is a general-purpose, high-level programming language with low-level facilities which supports both object-oriented and generic programming, popular in commercial computing since the 1990s (see [Information Technology Industry Council 2003] for an overview).

[3] See [Terras 2006] for further exposition of the procedures which would need to be adopted to undertake knowledge elicitation and investigate automated record linkage.

Works Cited

AHDS 2006 Arts and Humanities Data Service (AHDS). *Cross Search Catalogue*. http://www.ahds.ac.uk/catalogue/search.htm (2006). Accessed 31st October 2006.

AHDS History 2004 Arts and Humanities Data Service (AHDS) History. *Organising Waiver of Deposit*. http://ahds.ac.uk/history/depositing/waiver-of-deposit.htm (2004). Accessed 15th November 2006.

AHDS History 2005 Arts and Humanities Data Service (AHDS) History. *Invitation to Deposit*. http://ahds.ac.uk/history/depositing/invitation-to-deposit.htm (2005). Accessed 15th November 2006.

AHRC 2006 Arts and Humanities Research Council (AHRC), . *e-science. Background*. *AHRC ICT Methods Network Workgroup on Digital Tools Development for the Arts and Humanities*. 2006. http://www.ahrcict.rdg.ac.uk/activities/e-science/background.htm.

Adams 1995 Adams, John. *Risk*. New York & London: Routledge, 1995.

Anderson 2001 Anderson, Ross J. *Security Engineering*. Indianapolis: Wiley Publishing, 2001.

BBC 2004 BBC News. *Where Have All The Men Gone?* http://news.bbc.co.uk/1/hi/magazine/3601493.stm.

Blaikie 2005 Blaikie, Andrew, Eilidh Garrett and Ros Davies. "Migration, Living Strategies and Illegitimate Childbearing; A Comparison of Two Scottish Settings: 1871-1881". In Alysa Levene Samantha Williams and Thomas Nutt, eds., *Illegitimacy in Britain, 1700-1920*. London: Palgrave Macmillan, 2005.

Brodlie et. al. 2004 Brodlie, Ken, David Duce, Julian Gallop, Musbah Sagar, Jeremy Walton and Jason Wood. "Visualization in Grid Computing Environments". *Visualization* IEEE 2004 (2004), pp. 155-162.

Burnett 1980 Burnett, Claude A., Carl W. Tyler, Albert K. Schoenbucher and Jules S. Terry. "Use of Automated Record Linkage to Measure Patient Fertility after Family Planning Service". *American Journal of Public* 70: 3 (1980).

CamPop 2006 CamPop. *Determining the Demography of Victorian Scotland through Record Linkage* http://www-hpss.geog.cam.ac.uk/research/projects/victorianscotlanddemography/ (2006). Accessed November 3rd 2006.

Crocket et. al. 2006 Crocket, A., Jones, C. E., and Schrer, K. *The Victorian Panel Study*. Report Submitted to the ESRC (Award Ref: RES-500-25-5001) (2006).

Davies & Withers 2006 Davies, William, and Kay Withers. *Public Innovation: Intellectual Property in a Digital Age*. London: Institute for Public Policy Research, 2006.

Davies 2005 Davies, R., and E. Garrett. "More Irish Than the Irish? Nuptiality and Fertility Patterns on the Isle of Skye, Scotland 1881-1891". In Robert John Morris and Liam Kennedy, *Ireland and Scotland: Order and Disorder, 1600-2000*. Edinburgh: John Donald, 2005.

Dillon and Thorvaldsen 2001 Dillon, L.Y., and G. Thovaldsen. "A Look Into the Future Using and Improving International Microdata for Historical Research". In P.K. Hall R. McCaa and G. Thorvaldsen, eds., *Handbook of International Historical Microdata for Population Research*. Minneapolis: International Microdata Access Group, 2001. pp. 347-354.

Dunn & Blanke 2009 Dunn, Stuart, and Tobias Blanke. "Digital Humanities Quarterly Special Cluster on Arts and Humanities e-Science". *Digital Humanities Quarterly* 3: 4 (2009). http://www.digitalhumanities.org/dhq/vol/3/4/000079/000079.html.

Dunn 1946 Dunn, H.L. "Record Linkage". *American Journal of Public Heath* 36 (1946), pp. 1412-1416.

European Parliament 1996 European Parliament. *Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the Legal Protection of Databases*. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML (1996) Accessed 15th November 2006.

Fleury 1956 Fleury, M., and Henry, L. *Des Registres Paroissiaux l'Histoire de la Population. Manuel de Dpouillement et d'Exploitation de l'tat Civil Ancien*. Paris: Editions de l'INED (1956).

Fure 2000 Fure, Eli. "Interactive Record Linkage: The Cumulative Construction of Life Courses". *Demographic Research* 3: 11 (2000). http://www.demographic-research.org/Volumes/Vol3/11/html/0.htm.

Garrett 2003 Garrett, E., and R. Davies. "Birth Spacing and Infant Mortality on the Isle of Skye, Scotland, in the 1880s; A Comparison with the Town of Ipswich, England". *Local Population Studies* 71 (2003), pp. 53-74.

Garrett 2006a Garrett, Eilidih. "Urban-Rural Differences in Infant Mortality: a View from the Death Registers of Skye and Kilmarnock". In Eilidh Garrett Chris Galley, Nicola Shelton and Robert Woods, eds., *Infant Mortality: A Continuing Social Problem?* Ashgate, 2006. pp. 119-148.

Gutmann 1977 Gutmann, M.P. "Reconstituting Wandre. An Approach to Semi-Automatic Family Reconstitution". *Annales de Demographie Historique* (1977), pp. 315-341.

Higgs 2005 Higgs, E. *Making Sense of the Census Revisited: Census Records for England and Wales 1801-1901: A Handbook for Historical Researchers*. London: Institute of Historical Research, 2005.

History and Computing 1992 *History and Computing*. Special Issue on Record Linkage, 4.1.

History and Computing 1994 *History and Computing*. Special Issue on Record Linkage II, 6.3.

History and Computing 2006 *History and Computing*. Special Issue: *Longitudinal and Cross-Sectional Historical Data: Intersections and Opportunities*. 14.1/14.2.

Holmes 2006 Holmes, R. *The Accuracy and Consistency of the Census Returns for England 1841-1901 and their Indexes*. School of Library, Archive and Information Studies, University College London. M.A. Dissertation (2006).

Huntington et. al. 2007 Huntingdon, P., D. Nicholas and H.R. Jamali. "Employing Log Metrics to Evaluate Search Behavior and Success: Case Study, the BBC Search Engine". *Journal of Information Science* 33: 5 (2007), pp. 584-597.

ISO 2005 International Organization for Standardisation (ISO). "ISO/IEC 17799:2005 Information technology -- Security techniques -- Code of Practice for Information Security Management" Available from http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=39612ICS1=35ICS2=40ICS3 (2005) Accessed November 13th 2006.

Impedovo 1993 Impedovo, S. *Fundamentals in Handwriting Recognition*. Chateau de Bonas, France: Springer-Verlag, 1993.

Information Technology Industry Council 2003 Information Technology Industry Council. *Programming languages C++*. Second edition, Geneva: ISO/IEC. 14882:2003(E) (2003).

Inman 2002 Inman, Phillip. "Genealogy". *The Guardian* (September 26, 2002). http://www.guardian.co.uk/internetnews/story/0,,798781,00.html.

Katz 1972 Katz, M., and J. Tiller. "Record-Linkage for Everyman: A Semi-Automated Process". *Historical Methods Newsletter* 5 (1972), pp. 144-150.

Kingston Local History Project 2006 "The Kingston Local History Project" http://fass.kingston.ac.uk/research/local-history/projects/klhp/ (2006). Accessed 3rd November 2006.

Lloyd et. al. 2004 Lloyd, D., Webber, R., Longley, P. "Surnames as a Quantative Resource: The Geography of British and Anglophone Surnames". Conference, UCL, 28th-31th April, 2004, paper synopses available at http://www.casa.ucl.ac.uk/surnames/papers.htm (2004). Accessed 9th November 2006.

Mansfield 2006 Mansfield, M. "Ancestry Census Records: Background, Technology, Structures". Presentation for *ReACH All Hands Workshop*, UCL, 14th June 2006.

McGraw & Harbison-Briggs 1989 McGraw, Karen L., and Karen Harbison-Briggs. *Knowledge Acquisition: Principles and Guidelines*. London: Prentice-Hall International Editions, 1989.

McGraw & Westphal 1990 McGraw, Karen L., and Christopher Ralph Westphal. *Readings in Knowledge Acquisition: Current Practices and Trends*. London: Ellis Horwood Limited, 1990.

Mills et. al. 1989 Mills, D., Pearce, C. Davies, R., Bird, J., and Lee, C. *People and Places in the Victorian Census: A Review and Bibliography of Publications Based Substantially on the Manuscript Census Enumerators' Books 1841-1911*. Historical Geography Research Series, No. 23. Historical Geography Research Group of the Royal Geographical Society with the Institute of British Geographers (1989).

Newcombe 1988 Newcombe, H.B. *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford: Oxford University Press, 1988.

Nicholas et al 2006 Nicholas, D, P. Huntington, H.R. Jamali and C. Tenopir. "Finding Information in (Very Large) Digital Libraries: A Deep Log Approach to Determining Differences in Use According to Method of Access". *Journal of Academic Librarianship* 32: 2 (2006), pp. 119-126.

Nitsch 2006 Nitsch, D., S. Morton, B.L. DeStavola, H. Clark and D.A. Leon. "How Good is Probabilistic Record Linkage to Reconstruct Reproductive Histories? Results from the Aberdeen Children of the 1950s Study". *BMC Medical Research Methodology* 6 (2006).

PRDH 2000 Programme de Recherche en Demographie Historique. *The 1852 and 1881 Historical Censuses of Canada, 1881 Cleaning Manual, Phase 1* (2000). Available at http://www.prdh.umontreal.ca/census/en/uguide/OLD/1881projects.html Accessed 14/11/06.

Perkyns 1991 Perkyns, A. "Birthplace Accuracy in the Censuses of Six Kentish Parishes 1851-1881," in *Local Population Studies*, 47 (reprinted in Mills, D.R., and Schurer, K. (eds.) (1996), *Local Communities in the Victorian Census Enumerators' Books*, Oxford, Leopards Head Press.)

Perkyns 1993 Perkyns, A. "Age Checkability and Accuracy in the Censuses of Six Kentish Parishes 1851-1881," in *Local Population Studies*, 50 (reprinted in Mills, D.R., and Schurer, K. (eds.) (1996). *Local Communities in the Victorian Census Enumerators' Books* Oxford, Leopards Head Press.)

Reid 2006 Reid, A., R. Davies and E. Garrett. "Nineteenth Century Scottish Demography from Linked Censuses and Civil Registers: A 'Sets of Related Individuals' Approach". *History and Computing* 14: 1 (2006).

Riedel et. al. 2007 Riedel, M., T. Eickermann, S. Habbinga, W. Frings, P. Gibbon, D. Mallmann, F. Wolf, A. Streit, T. Lippert, W. Schiffmann, A. Ernst, R. Spurzem and W.E. Nagel. "Computational Steering and Online Visualization of Scientific Applications on Large-Scale HPC Systems within e-Science Infrastructures". Presented at *IEEE 2007. Proceedings of the IEEE International Conference on e-Science and Grid Computing* (2007), pp. 483-490.

Robey 2006 Robey, D. "AHRC-EPSRC-JISC Arts and Humanities e-Science Initiative, Research Grants and Studentships Scheme" Introductory Presentation, *e-Science Research Grants and Studentships Open Meeting*, 8 September 2006, Woburn House, 20 Tavistock Square, London (2006).

Sauleau 2005 Sauleau, E.A., J-P Paumier and A. Buemi. "Medical Record Linkage in Health Information Systems by Approximate String Matching and Clustering". *BMC Medical Informatics and Decision Making* 5: 32 (2005).

Schürer & Woollard 2002 Schürer, K. and Woollard, M. "National Sample from the 1881 Census of Great Britain 5% Random Sample. Working Documentation v1.1", University of Essex, Historical Census and Social Surveys Research Group. Available at http://www.data-archive.ac.uk/doc/4177%5cmrdoc%5cpdf%5cguide.pdf (2002). Accessed 9th November 2006.

Shadbolt & Burton 1990 Shadbolt, N., and M.A. Burton. "Knowledge Elicitation Techniques - Some Experimental Results". In Karen L. McGraw and Christopher R. Westphal, eds., *Readings in Knowledge Acquisition, Current Practices and Trends*. London: Ellis Horwood Limited, 1990.

Stallings 2005 Stallings, W. *Network Security Essentials: Applications and Standards*, Prentice Hall (2005).

Stallings 2008 Stallings, W. and Brown, L. *Computer Security: Principles and Practice*, Prentice Hall (2008).

Terras 2006 Warning: Biblio formatting not applied. MelissaTerras. *he Researching e-Science Analysis of Census Holdings Project: Final Report to AHRC*. 2006. www.ucl.ac.uk/reach/. AHRC e-Science Workshop scheme.

Tilley 2003a Tilley, P. "The Kingston Local History Project. Creating Life Histories and Family Trees for Communities in Victorian Britain". *IMAG Workshop Paper, Longitudinal and Cross-sectional Historical Data, Intersections and Opportunities*, Montreal, 10th and 11th November 2003.

Tilley 2003b Tilley, P. "A Restless Community. Preliminary Findings from a Study of Migration from Kingston on Thames in 1871". Paper presented to the PhD workshop, Economic History Department, London School of Economics 5th November 2003.

Tillot 1972 Tillot, P.M. "Sources of Inaccuracy in the 1851 and 1861 Censuses". In E.A. Wrigley, *Nineteenth-Century*

Society: Essays in the Use of Quantitative Methods for the Study of Social Data. Cambridge: Cambridge University Press, 1972. pp. 82-133.

UCL Research Computing 2006a UCL Research Computing, "Altix". http://www.ucl.ac.uk/research-computing/services/altix/index.html (2006a). Accessed 13th November 2006.

UCL Research Computing 2006b UCL Research Computing "Services" http://www.ucl.ac.uk/research-computing/information/services (2006b). Accessed 13th November 2006.

Warwick et al. 2008 Warwick, C., M. Terras, P. Huntington and N. Pappa. "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 (2008), pp. 85-102.

Winchester 1970 Winchester, I. "The Linkage of Historical Records by Man and Computer: Techniques and Problems". *Journal of Interdisciplinary History* 1: 1 (1970), pp. 107-124.

Winkler 2001 Winkler, W. E. "Records Linkage Software and Methods of Merging Administrative Lists". *Bureau of the Census Statistical Research Division, Statistical Research Report Series RR2001/3* (2001). Available at http://www.census.gov/srd/papers/pdf/rr2001-03.pdf. Accessed November 16th 2005.

Woollard 1997 Woollard, M. "'Shooting the Nets': a Note on the Reliability of the 1881 Census Enumerators Books". *Local Population Studies* 59 (1997), pp. 54-57.