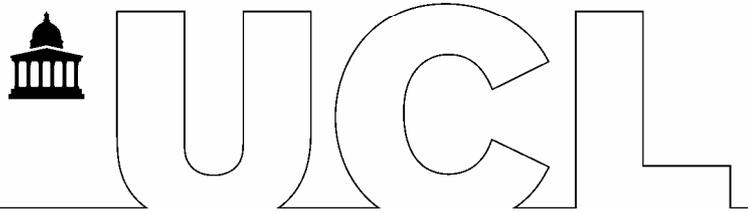


UCL ReACH: RESEARCHING e-SCIENCE ANALYSIS OF CENSUS HOLDINGS
SCHOOL OF LIBRARY, ARCHIVE AND INFORMATION STUDIES
REPORT TO AHRC



ReACH: Researching e-Science Analysis of Census Holdings
AHRC Arts and Humanities e-Science Workshop Series

ReACH

<http://www.ucl.ac.uk/reach/>

Project Report

Dr Melissa Terras, UCL SLAIS

V.1.2 06/12/06

ReACH Workshop Series: Executive Summary

e-Science has 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 datasets, and little understanding of or access to e-Science technologies. The ReACH 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.

The ReACH project, based in the School of Library, Archive, and Information Studies at UCL¹, worked in close collaboration with:

- The National Archives², 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)
- Ancestry.co.uk³, 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
- UCL Research Computing⁴, the UK's Centre for Excellence in networked computing.

ReACH consisted of three one day workshops held at UCL in Summer 2006, bringing together expertise across different domains to ascertain how useful, possible, or feasible undertaking e-Science analysis of historical census material would be.

¹ <http://www.ucl.ac.uk/slais/>

² <http://www.nationalarchives.gov.uk/>

³ <http://www.ancestry.co.uk>

Experts in history, archives, genealogy, computing science, and humanities computing also contributed to the invitation only workshop series. The workshops covered the academic, technical, and legal issues in setting up a pilot project which would aim to analyse datasets from Ancestry and The National Archives using the high performance computing facilities available at UCL.

The project aimed to: highlight issues regarding the application of e-Science technologies to Humanities datasets, develop a project proposal for full scale analysis of Ancestry.co.uk's historical datasets utilising Research Computing facilities at UCL, bring together a wide range of interdisciplinary expertise to ensure best practice, highlight any issues of concern which would preclude a large scale project from being useful or successful, ascertain the historians viewpoint of the benefits and concerns in undertaking a larger scale project, predict the form and type of results which would emanate from a future project with the available datasets, and ascertain the comprehensiveness and accuracy of the available transcribed datasets.

The results of the well attended workshop series was a sketch for a potential project, and recommendations regarding the implementation of e-science (high performance computing) technologies in this area. However, at this time, it was not thought possible to pursue the potential project in the following e-Science call from the AHRC due to a variety of reasons which are elucidated in this report. As a result of the workshop series, the ReACH workshop series proposes the following recommendations:

For the historian:

⁴ <http://www.ucl.ac.uk/research-computing/>

- 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. This will change as more data is digitised (and becomes available to the general research and genealogical community through publicly available websites).
- The potential for high performance processing of large scale census data is large, and may result in useful techniques and datasets (for both historian and genealogist) when adequate census data becomes available. This should be revisited in the future. Access to computational techniques or expertise or managerial issues are not the limiting factors here.

For researchers in e-Science and the Arts and Humanities:

- High performance computing and e-Science community 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.
- Often the problems facing e-Science research in the arts and humanities are not technical. Although there is still fear in using high performance computing 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.
- Using the processing power of computational grids may be unnecessary if 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 queuing jobs on a stand alone high performance machine requires less managing at present than using processing power dispersed over a local, national, or international grid.
- Finding arts and humanities data which is of a large enough size to warrant grid (or high performance computing) processing whilst being of high enough quality can be a problem for a researcher wishing to high performance computing 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. 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.

- The high performance computing facilities at University College London are available for research in the humanities – and there is potential here for providing the computational facilities for projects such as the Victorian Panel Study (Crocket, Jones and Schürer (2006)).
- 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. Although not necessary to include these in a funding bid, it is important to ascertain that the project has access to infrastructures which will allow it to negotiate licenses and contracts. The importance of this issue cannot be stressed enough – especially when the project is wholly dependant 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.
- 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 with 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 large when it comes to data management, IPR, copyright, and licensing issues.

- 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.

For funding councils:

- Where arts and humanities e-Science 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.
- e-Science projects in the arts and humanities may also 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 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. High performance computing is as much a part of “e-Science” in the sciences as distributed 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.

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.

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 Wolfram (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.

Contents

| | |
|--|------|
| Executive Summary | i |
| Acknowledgements | viii |
| Contents | xi |
| 1. Introduction | 1 |
| 2. Methodology | 4 |
| 2.1. Workshops | 4 |
| 2.2. Follow up to workshops | 6 |
| 3. Findings | 7 |
| 3.1. Benefits to Historians | 7 |
| 3.2. Technical Implementation | 15 |
| 3.3. Managerial Issues | 17 |
| 3.4. Future Research | 23 |
| 3.4.1. Knowledge Elicitation and the Historian | 29 |
| 3.4.2. Automated Record Linkage | 32 |
| 3.5. Taking the Project Forward | 35 |
| 4. Recommendations | 37 |
| 5. Conclusion | 42 |
| 6. References | 44 |
| Appendix A | 52 |
| A.1 Workshop 1: All Hands Workshop | 53 |
| A.1.1 Programme | 53 |
| A.1.2 Detail of Census Fields Available | 59 |
| A.1.3 Attendees | 61 |
| A.2 Workshop 2: Technical Workshop | 62 |
| A.2.1 Programme | 62 |
| A.2.2 Attendees | 63 |
| A.3 Workshop 3: Managerial Workshop | 64 |
| A.3.1 Programme | 64 |
| A.3.2 Attendees | 65 |
| A.4 Steering Group | 65 |
| A.4.1 Programme | 65 |
| A.4.2 Committee | 68 |

1. Introduction

e-Science technologies have the potential to enable large-scale datasets to be searched, analysed, and shared quickly, efficiently, and in complex and novel ways. So far, little application has been made of the processing power of grid technologies to humanities data, due to lack of available datasets, and little understanding of or access to e-Science technologies. The ReACH workshop series was established to investigate the potential application of e-Science and high performance computing technologies to a large dataset of interest to historians, humanists, digital consumers, and the general public: historical census records.

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 (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 Ancestry⁵, Genes Re-united⁶, QinetiQ⁷, British Origins⁸, The Genealogist⁹, and 1837Online¹⁰) have digitised vast swathes of historical census material (although to varying degrees of completeness and accuracy). There is much

⁵ <http://www.ancestry.com/>

⁶ <http://www.genesreunited.co.uk/>

⁷ <http://www.qinetiq.com/>

⁸ <http://www.origins.net/BOWelcome.aspx>

⁹ <http://www.thegenealogist.co.uk/>

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, Jones and Schürer (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 Project¹¹). However, there have been few opportunities for the application of high performance computing to utilise large scale 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).

The aim of the ReACH series was to bring together disparate expertise in Computing 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 Studies¹², who have expertise in digital humanities and advanced computational techniques, as well as digital records management,
- The National Archives¹³, 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),

¹⁰ <http://www.1837online.com/>

¹¹ <http://www.nappdata.org/napp/>

¹² <http://www.ucl.ac.uk/slais/>

¹³ <http://www.nationarchives.gov.uk>

- Ancestry.co.uk¹⁴, 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. The input of Ancestry was central to this research to gain access to the complete range of UK census years in digital format,
- UCL Research Computing¹⁵, the UK's Centre for Excellence in networked computing, who have extensive high performance computing facilities available for use in research.

The 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. 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 for future projects. The project also wanted to establish how feasible, and indeed useful, undertaking such an analysis of historical census data would be. The results of the well attended workshop series was a sketch for a potential project, and recommendations regarding the implementation of e-science (high performance computing) technologies in this area. However, at this time, it was not thought possible to pursue the potential project in the following e-Science call from the AHRC due to a variety of reasons which are elucidated in this report.

The following report describes the methodology of the workshops, reporting on suggestions made during the day, sketching out a future project regarding how historical census material can be analysed utilising high performance computing, and

¹⁴ <http://www.ancestry.co.uk>

¹⁵ <http://www.ucl.ac.uk/research-computing/>

extrapolates recommendations that can be applied in general to the use of e-Science in the arts and humanities research sectors. Appendices are provided regarding materials handed out at the workshops, and attendees of the sessions¹⁶.

2. Methodology

2.1 Workshops

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, throughout summer 2006. The workshops were split into three topics:

- The All Hands Workshop (Wednesday 14th June 2006) 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. A full programme and list of attendees is available in Appendix A.1.

¹⁶ All individuals named as attending and presenting at the workshops and providing other information to the project gave their permission to be included in this report.

- The Technical workshop (Thursday 15th June 2006) 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 provided the technical information necessary for implementation of a full scale project. 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 (access to existing kit), software required, software development issues, security issues any other technical issues which may arise. A full programme and list of attendees is available in Appendix A.2.
- The Managerial Workshop (Tuesday 25th July 2006) 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 of a full scale project, financial structure of a full scale project, paths to dissemination and publicity, and other topics suggested by participants. A full programme and list of attendees is available in Appendix A.3.

Minutes were taken throughout each workshop, and discussion sessions were recorded (though not transcribed) to allow clarification of discursive elements.

2.2 Follow Up to Workshops

Following the workshops, points of interest raised 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 at Kingston University London¹⁷ 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 Register¹⁸ 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 AstroGrid¹⁹ project (who were interested in seeing how results of a potential project could be useful for research involving scientific data).

¹⁷ <http://localhistory.kingston.ac.uk/contributePages/klhp.html>

¹⁸ <http://www.freebmd.org.uk/>

¹⁹ <http://www2.astrogrid.org/>

3. 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.

3.1 Benefits for Historians – Would this be useful?

There is significant interest in how high performance computing 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, 1994, 2006), PRDH (2000), Dillon and Thorvaldsen (2001), Schürer and 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:

- 1) Generate automatic matches of records throughout the census years available, creating what is known as a “longitudinal database” of individuals across the census. This will require: 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 and families and track population change across time, and inform other projects interested in building such datasets.
- 2) Generate rich variant lists for users. The use of variants is important in dealing with the problematic nature of census data, which can often have errors due to its nature of collection (see 4.1). By building up lists of common variants present in the UK census data, this will both 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 probro), 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.
- 3) Log analysis of usage statistics from those accessing historical census data online could be undertaken to see how users link data, and analyse different records. This could be useful to understand the nature of genealogical research, and also the procedures used to match records.

(see Nicholas *et al* 2006 and Huntingdon *et al* (forthcoming 2007) regarding how these techniques have been employed to understand digital user behaviour of other online resources, and Warwick *et al* (forthcoming 2007) for this technique applied particularly to those in the arts and humanities.)

- 4) The checking and cleansing of census data. The 5% sample of 1881 census data digitised and developed by Kevin Schürer and Matthew 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 and Woollard 2002 p.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” (p.16) 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 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 and Woollard 2002 Appendix C for a detailed discussion of the procedures carried out.)
- 5) Develop 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 used).

- 6) The generation of techniques 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.
- 7) Name mapping of geography to names. There has been some success with this – a UCL project based in the Centre for Advanced Spatial Analysis²⁰ has been working on a Surname Profiler²¹ which investigates the distribution of surnames in the UK in both historic (1881) and contemporary (1998) census datasets. (A conference was held at UCL regarding the benefits this has for research between 28th-31st April 2004, see Lloyd, Webber, and Longley (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.
- 8) 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

²⁰ <http://www.casa.ucl.ac.uk/>

population reports (which are now being digitised themselves by the Online Historical Population Reports Project²²) 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, through 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.

- 9) Assign 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).
- 10) Adding current geographical data to the census. Similar to 9, above, it is a common request at the National Archives for people to be able to

²¹ <http://www.spatial-literacy.org/UCLnames/>

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.

- 11) Visualisation techniques can be used 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 the geo-spatial data (see 9 and 10, above) is available, it can be manipulated through GIS, increasing the means to interrogate and conduct new research with the data.
- 12) Calculating and identifying individuals who have been missed out in various censuses. 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.

²² <http://www.histpop.org/demo-b/servlet/Show?page=Home>

13) Missing data in the digital records can be reconstituted through contextual information (for example, street numbers are missed out 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 (providing that original transcripts are maintained and data integrity preserved for quality control, as in the enriched dataset in Schürer and Woollard (2002)).

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 (Mansfield 2006), including image files. 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/index.php> for an overview), making use of the high performance computing facilities at UCL would allow this data to be interrogated in a reasonable and realistic timeframe.

However, whether using high performance computing 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

e-Science thus stands for 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 ICT 2006)

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

‘**e-Science**’ stands for 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 3.2). 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.

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 3.3). Therefore, distribution of the *results* of the project may not be possible via Internet or Grid.

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.

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 in 4.1. Managerial issues regarding data security and procedures are covered in 3.3.

3.2 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/services/>, but can be summarised as

- AccessGrid facilities for virtual collaboration.
- Central Computing Cluster (C³) 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.
- The Altix for High-Performance Computing. (UCL Research Computing 2006b)

For the security reasons outlined in 3.1 and 3.3, any project would have to use a stand alone machine rather than distribute data via a network (such as the Condor computing pool) for processing via a/the grid. After consultation with UCL Research Computing regarding memory requirements, scalability and I/O profile, it was determined that the SGI Altix²³ 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 2006). 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++²⁴ based programmes: a standard in the development of software tools. C++ routines could be

²³ See <http://www.sgi.com/products/servers/altix/> for more information about Altix machines.

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. Of course, the difficulty lies in *what* to program. This is covered in 3.4.2.

Because of security issues, data would be received from Ancestry on 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)²⁵. 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)²⁶, which looks at, amongst other things, security and privacy of clinical data. Recommendations regarding security procedures are made in the following section.

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

²⁴ 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).

²⁵ <http://www.chime.ucl.ac.uk/>

²⁶ <http://www.ucl.ac.uk/research-computing/research/e-science/clef.html>

technical barriers to proceeding with this manner, and the facilities at UCL are even available free of charge for research²⁷.

3.3 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.

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 is 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.

Security issues regarding dealing with commercially sensitive data need to be resolved before delivery of the data is taken. Consultation with data management expertise in CHIME resulted in the recommendation of ISO/IEC 7799:2005, a

²⁷ Under Full Economic Costing, the Altix machine costs 44 pence per CPU hour (UCL Research Computing 2006b). An indepth analysis of algorithms needed would have to be undertaken to ascertain how many CPU hours would be required before costing this part of the project.

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. The standard covers the following areas:

- security policy;
- organization of information security;
- asset management;
- human resources security;
- physical and environmental security;
- communications and operations management;
- access control;
- information systems acquisition, development and maintenance;
- information security incident management;
- business continuity management;
- compliance.

and most importantly, 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 providing data in advance 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 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) and Stallings (2005, forthcoming 2007).

Legal agreements should also be undertaken about the fair use and application of the data for the duration of the project, and what happens to the data after the project ends. This will require assistance from institutional lawyers (who often provide the service free to the project on behalf of the institution, so 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.

There are also considerations that need to be made regarding what happens to the data at the end of the project, or where the results are kept. Firstly, it is a condition of AHRC funding to offer any resulting digital output for deposit with AHDS to ensure its long term viability and dissemination (see AHDS History 2005), which obviously poses a problem for data which has been supplied from a commercial partner which they do not wish to be made available in the public domain. However, it is possible to negotiate a waiver of deposit with the AHRC and AHDS (see AHDS History 2004). Although the examples given for reasons why a waiver may be granted do not include use of sensitive commercial data, when consulted with, AHDS History agreed that this could be a reason for negotiating a waiver.

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 particularly 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 and 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 the 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 the project, to enable researchers to use and publish results, protect the commercial rights of the company, but also protect the intellectual investment of the researchers, especially regarding any outcomes which may be suitable for knowledge transfer or technological spin-off.

In a case such as this, 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 may 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.

For a grant application, it is important to establish 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
- UCL Business PLC, and the Contract Research Office, will arrange IPR agreements and commercial exploitation
- the foreground IP of the project will remain the property of UCL
- Commercial background IPR (data, etc) will be licensed accordingly
- UCL Business and the related infrastructure can assist in all of this
- 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.

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.

3.4 Future Research

Following consultation with historians, it was obvious that the most popular, useful, and likely to succeed, project which could emanate 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 high performance computing facilities (meaning that computational time would not be of concern to the project) it would 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 high performance computing across fuzzy datasets
- develop tools for historians undertaking the construction of longitudinal datasets, to aid them in checking and investigating possible linkages across datasets.

The knowledge transfer opportunities from developing robust and benchmarkable techniques would be large: consultation with Physicists working on the AstroGrid²⁸, for example, revealed that they are facing the same problem: being able to track and trace individual entities across fuzzy and incomplete datasets. Data sets 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 a 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).

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), Perkyms (1991, 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 and Woollard 2002), census data will

²⁸ <http://www2.astrogrid.org/>

remain “fuzzy”, and often incomplete. This makes computational matching of data difficult.

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 out 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.

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 FreeBMD²⁹ 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 under way, 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

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 2006³⁰)

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.

The Kingston Local History³¹ group are also interested in linking records across the different census years, and are 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:

- Census Enumerators Books for Kingston Upon Thames Census Area - 1851 to 1891
- Bonner Hill Cemetery Burial Registers - 1855 to 1911
- Kingston Parish Burial Registers - 1850 to 1901
- Kingston Parish Marriage Registers - 1850 to 1901

This obviously required much digitisation 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

²⁹ <http://www.freebmd.org.uk/>

³⁰ See also Garrett and Davies (2003), Davies and Garrett (2005) Blaikie, Garrett and Davies (2005), Reid, Davies, and Garrett (forthcoming 2006) and Garrett (forthcoming 2006) for research emanating from these datasets.

³¹ <http://localhistory.kingston.ac.uk/contributePages/klhp.html>

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.

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, 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, Jones, and Schürer, (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,
- Projects are dependent on the creation of more datasets (Campop, Kingston, and VPS.)

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.

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.

How would such a project proceed? A process of knowledge elicitation 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). 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.

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.

Some of the issues regarding work which would have to be undertaken regarding legal arrangements, and technical set up, have already been covered, above. This final section will cover two aspects of the project which would have to be addressed: understanding how historians undertake manual record linkage, and establishing and reviewing previous research into automated record linkage.

3.4.1 Knowledge Elicitation and the Historian

To be able to construct a system which can carry out record linkage as effectively as a human researcher it is necessary to understand how historical experts link data when they undertake this task manually, in order to make explicit the routines used, which can then be expressed computationally. The problem with trying to discover the techniques that historians use whilst undertaking record linkage is that experts are notoriously bad at describing what they are expert at (McGraw and Harbison-Briggs

1989). Experts utilise and develop many skills which become automated and so they are increasingly unable to explain their behaviour, resulting in the troublesome “knowledge engineering paradox”: the more competent domain experts become, the less able they are to describe the knowledge they use to solve problems (Waterman 1986). Added to this problem is the fact that, although techniques such as Knowledge Acquisition (conventionally defined as the gathering of information from any source) and Knowledge Elicitation (the subtask of gathering knowledge from a domain expert, see Shadbolt and Burton (1990)) are becoming increasingly necessary for the development of computer systems, there is no consensus within the field as to the best way to proceed in undertaking such research.

Discussions regarding how best to elicit knowledge for the basis of an expert system first appeared in the late 1970s (Feigenbaum 1977). Early attempts at eliciting, formalising, and refining expert knowledge were so unfruitful that Knowledge Elicitation was labelled the “bottleneck” to building knowledge-based systems (Feigenbaum 1977). Throughout the 1980s and early 1990s, protocols (often referred to as the “traditional” or “transfer” approach to Knowledge Elicitation) were developed regarding how a knowledge engineer should interact with a domain expert to organise and formalise extracted knowledge so that it is suitable for processing by a knowledge based system (Diaper 1989; McGraw and Harbison-Briggs 1989; Boose and Gaines 1990; McGraw and Westphal 1990; Morik, Wrobel *et al.* 1993). These discussions centre on the suitability of different techniques (often derived from clinical psychology and qualitative research methods used in the social sciences) include: unstructured, semi-structured, and focussed interviews with the expert(s), Think Aloud Protocols (TAPs), where an expert is set a task and asked to describe

their actions and thought processes, stage by stage; sorting, where the expert is asked to express the relationship between a pre-selected set of concepts in the domain; and laddering, where the expert is asked to explain the hierarchical nature of concepts within the domain. A discussion contrasting these with other knowledge acquisition techniques, highlighting their suitability regarding the elicitation of certain types of knowledge, can be found in Cordingley (1989). See Terras (2005 and 2006) for an example of how these techniques have previously been used with humanities experts – in that case, understanding how historians read ancient texts. Automated Knowledge Elicitation is also a possibility, although has little success with complex problems³².

To be able to construct a computational system that would replicate the techniques historians use for record linkage across a broad scale of material, an in depth Knowledge Library would have to be constructed (gathering all available research regarding record linkage across the census) and a program of Knowledge Elicitation carried out with experts in this area. This depends on having access to experts willing to work with a researcher on this task. The task is also not something that can be

³² Since the early 1990s the field has focussed on Automated Knowledge Elicitation, incorporating the same psychological techniques into computer programs, to make the interactions more productive, assisting, and in some cases replacing, the knowledge engineer (White 2000). Such tools were, at first, implemented in a stand-alone, domain independent way, focussing on the collection of particular types of data. For example, the ETS and the AQUINAS systems (Boose 1990) are computerised representations of the Repertory Grid method (see Kelly 1955) and have been used to derive “hundreds” of small knowledge-based systems. Gradually, programs appeared offering implementations of various techniques bundled together as a Knowledge Elicitation “workbench” such as the research prototype ProtoKEW (Reichgelt and Shadbolt 1992) which was later repackaged and marketed as the commercial PC-Pack system (<http://www.epistemics.co.uk/Notes/55-0-0.htm>) Researchers have now started to utilise the internet, developing distributed knowledge acquisition tools such as RepGrid³² (<http://repgrid.com/>) and WebGrid, (<http://pages.cpsc.ucalgary.ca/~gaines/WebGrid/>) which can be used remotely to build up knowledge bases and data sets. However, these computer tools produce the best results when applied to very small domains to build knowledge based systems which carry out well-defined tasks, and are not successful at providing overviews of complex systems, or when used to describe domains about which little is known from the outset (Marcus 1988; White 2000).

rushed: a proper investigation would involve a knowledge engineer working with three or four experts over the course of at least a year (but probably two), to be able to analyse, distil, and express the techniques used in such a way as could be undertaken computationally.

3.4.2 Automated Record Linkage

Further to building up an understanding regarding how experts match records manually, it will also be necessary to undertake a comprehensive review of Automated Record Linkage techniques which have been used for textual data (both applied to historical records and to automated records linkage attempts in other domains). Only by exploring, reviewing and understanding the known literature and research on this topic will it be possible to ascertain how these techniques can be applied to matching historical census data, and if other techniques exist which have not yet been applied to census data, which may be useful to this domain.

The available literature on automatic record linkage is large and expansive. There has been much research into the use of automated record linkage since the early days of computer processing (Dunn 1946), for many applications. The linkage of records which refer to the same entity in separate data collections without a unique database key is a common requirement in public health (for a standard text on the subject see Newcombe (1988), and for just a few examples, see Burnett *et al* 1980, Nitsch *et al* 2006), government administration (Winkler (2001)), historical research (History and

Computing 2006) and biomedical research (Sauleau *et al* 2005). Automatic record linkage has obvious advantages: once the source material is digitized and the system is developed, record linkage can be accomplished very quickly. Algorithmic expression of matching makes the linkage consistent and transparent: results should be reproducible, and the automated process is able to handle larger volumes of data than human researchers can.

The historical community has been attempting to create record linkage techniques since the 1950s. The “Family reconstruction” or “reconstitution technique” was a time consuming manual process developed by Fleury and Henry (1956), developed for linking small French parishes. This encountered problems when applied data covering larger areas, or dealing with populations in which there is much social mobility. Throughout the 1970s there were several attempts to undertake family reconstitution computationally, by either wholly- or semi- automated record linkage, to reconstruct both families and individual life cycles (Winchester (1970), Katz and Tiller (1972), Gutmann (1977)). Projects were established, many of which are still undertaking research today: for example the Le Programme de Recherche en Démographie Historique at the University of Montreal³³. Specific semi-automated techniques are still being developed (Fure (2000), Tilley (2003a)). However, programs developed so far tend to be specific to certain datasets, and are not scaleable without comprehensive adaptation. Linking algorithms tend to create more erroneous links than corresponding manual procedures. Additionally, fully automated systems find non-systematic errors in the sources problematic (see Fure (2000) for an overview). Additionally, there is often no way of checking or benchmarking these routines to

³³ <http://www.genealogie.umontreal.ca/>

ascertain how effective they have been over anything but a small manually constructed dataset. Finally, construction of these systems can be a time consuming and costly endeavour.

There is not room here to explore the mathematical underpinnings of techniques which are used to carry out Automated Record Linkage, but for a project to enter this vast and widespread domain, it will be necessary to understand prior research, whilst being able to mobilise techniques used to formalise knowledge and express uncertainty, undertaking a combination of probabilistic and rule based approaches to develop routines which can deal with fuzzy and complex datasets such as that contained within the census records. This is not a project to undertake lightly: 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. The project would require at least two to three years of research and development for a researcher to undertake analysis of available literature and explore techniques which could be applied to the census material.

Unfortunately, should a record linkage project be carried out on the Kingston Upon Thames area, developing routines which could be checked against the database which has already been constructed and checked by researchers, at current time of writing data is 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.

3.5 Taking the Project Forward

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 the project would not apply for funding in the AHRC/EPSRC e-Science call for a variety of reasons.

- The funding available, once the project was costed, would not be enough to cover a three year project (which is the minimum duration a useful project would have to be).
- The research is high risk - it would take at least two years of development before anything could be demonstrated to the historical community (and even then, it may not be of use to the historical community).
- The bid would fall down on the Arts and Humanities versus Economic and Social Research divide: the analysis of historical census material is of interest in both areas, but who should fund this research? This is a common problem faced by researchers working on the census material.
- Although Kingston, Free BMD, and Ancestry, have pledged to give data, with results being tested on the Kingston upon Thames area, it is unlikely that any algorithms developed would be scaleable at this time - there is not comprehensive enough data to scale the results of the research up to any other other place in England, for various reasons of quality and quantity of available data. (Scotland may be another matter). This project should be

revisited once suitable data to enable scaling up the datasets becomes available.

- There were also problems with the scope and the quality of some of the data available, and it was felt by the historians that this would not benefit them yet. The project was advised to revisit this in a few years time when larger datasets of higher quality may be available.
- Due to the commercially sensitive nature of the datasets we would not be able to deposit the datasets in the public domain, nor make them available over the National Grid Service, or Internet. Given the security measures required, the project would use the stand-alone High Performance computing facilities at UCL. This may exclude the project from being “e-Science”.

4. Recommendations

Although it has been decided not to take the project forward at this stage, until cleaner, more extensive data, and larger funding streams may be available, the ReACH series has resulted in various recommendations.

For the historian:

- 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).
- 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 are not the limiting factors here.

For researchers in e-Science and the Arts and Humanities:

- High performance computing and the e-Science community 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.
- Often the problems facing e-Science research in the arts and humanities are not technical. Although there is still fear in using high performance computing 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.
- Using the processing power of computational grids may be unnecessary if 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 queuing jobs on a

stand alone high performance machine requires less managing at present than using processing power dispersed over a local, national, or international grid.

- 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 high performance computing 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.
- The high performance computing facilities at University College London are available for research in the humanities – and there is potential here for providing the computational facilities for projects such as the Victorian Panel Study (Crocket, Jones and Schürer (2006)).
- 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. Although it is not necessary to include details of these in a funding bid, it is important to ascertain that the project has access to infrastructures which will allow it to negotiate licenses and contracts. The importance of this issue cannot be stressed enough – especially when the project is wholly dependant 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.
- 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 large when it comes to data management, IPR, copyright, and licensing issues.
- 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.

For Funding councils:

- Where e-Science 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.

- e-Science projects in the arts and humanities may also 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 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. High performance computing 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 focusses 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.

5. 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.

There was much interest in the series, as the topic of how high performance computing 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?

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 high performance computing in this area if and when comprehensive data sets of high enough quality become available.

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 procedures which 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 outwith the arts and humanities domain. Tackling e-Science projects in the arts and humanities may then inform developments in computer science for other applications.

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.

6. References

Adams, J. (1995) Risk. Routledge.

Anderson, R. (2001). Security Engineering. John Wiley and Sons.

AHDS History (2004). Organising Waiver of Deposit.
<http://ahds.ac.uk/history/depositing/waiver-of-deposit.htm> Accessed 15th November 2006.

AHDS History (2005). Invitation to Deposit. <http://ahds.ac.uk/history/depositing/invitation-to-deposit.htm> Accessed 15th November 2006.

Arts and Humanities Data Service (AHDS) (2006), Cross Search Catalogue.
<http://www.ahds.ac.uk/catalogue/search.htm?q=n&q=census&s=all&coll=y&item=y>.
Accessed 31st October 2006.

Arts and Humanities Research Council (AHRC) (2006) “e-Science, Background”. AHRC ICT Programme Activities and Services, <http://www.ahrcict.rdg.ac.uk/activities/e-science/background.htm> Accessed 13th November 2006.

BBC (2004). “Where have all the men gone?” Magazine.
<http://news.bbc.co.uk/1/hi/magazine/3601493.stm> Accessed 13th November 2006.

Blaikie, A. Garrett, E. and Davies, R. (2005). “Migration, living strategies and illegitimate childbearing; a comparison of two Scottish settings: 1871-1881”, in A. Levene, T. Nutt and S. Williams eds. (2005). Illegitimacy in Britain, 1700-1920, Palgrave Macmillan, 141-167.

Boose, J. H. (1990). "Uses of Repertory Grid-Centered Knowledge Acquisition Tools for Knowledge-Based Systems." In J. Boose and B. Gaines (1990) Foundations of Knowledge Acquisition. London, Academic Press: 61-83.

Boose, J. and Gaines, B. (eds). (1990). The Foundation of Knowledge Acquisition. London, Academic Press.

Burnett, C.A, Tyler, C. W., Schoenbucher, A. K. and Terry, S. J. (1980) “Use of automated record linkage to measure patient fertility after family planning service” American Journal of Public Health. 1980 March; 70(3): 246–250.

CamPop 2006, “Determining the Demography of Victorian Scotland through Record Linkage” <http://www-hpss.geog.cam.ac.uk/research/projects/victorianscotlanddemography/>. Accessed November 3rd 2006.

Cordingley, E. (1989). "Knowledge Elicitation Techniques for Knowledge Based Systems." In Diaper, D., Ed. (1989). Knowledge Elicitation: Principles, Techniques and Applications. Chichester, Ellis Horwood: 90-103.

Crocket, A., Jones, C. E., and Schürer, K. (2006). “The Victorian Panel Study”. Report Submitted to the ESRC (Award Ref: RES-500-25-5001), May 2006.

Davies, R. and Garrett, E. (2005). “More Irish than the Irish? Nuptiality and fertility patterns on the Isle of Skye, Scotland 1881-1891”, in L. Kennedy & R. J. Morris, eds. (2005) Ireland and Scotland: Order and disorder, 1600-2000, Edinburgh.

Davies, W. and Withers, K. (2006). Public Innovation: Intellectual property in a Digital Age. Institute for Public Policy Research.

Dillon, L.Y. and Thorvaldsen, G. (2001). “A Look into the Future — Using and Improving International Microdata for Historical Research”. In Hall, P.K., McCaa, R. and Thorvaldsen, G. (eds). Handbook of International Historical Microdata for Population Research. International Microdata Access Group, Minnesota Population Center, Minneapolis, Minnesota. 347-354.

Diaper, D. (1989). Knowledge Elicitation: Principles, Techniques and Applications. Chichester, Ellis Horwood.

Dunn, H. L. (1946). "Record Linkage". American Journal of Public Health, Vol. 36, 1412-1416.

Feigenbaum, E. A. (1977). The Art of Artificial Intelligence: Themes and Case Studies in Knowledge Engineering. International Joint Conference of Artificial Intelligence (5, 1977): 1014-1029.

European Parliament (1996). 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> Accessed 15th November 2006.

Fleury, M., and L. Henry. (1956). Des registres paroissiaux à l'histoire de la population. Manuel de dépouillement et d'exploitation de l'état civil ancien. Paris: Editions de l'INED.

Fure, E. (2000). Interactive Record Linkage: The Cumulative Construction of Life Courses. Demographic Research, 3:11. <http://www.demographic-research.org/Volumes/Vol3/11/html/0.htm> Accessed 16th November 2006.

Garrett, E. (forthcoming 2006). "Urban-rural differences in infant mortality: a view from the death registers of Skye and Kilmarnock", in E. Garrett, C. Galley, N. Shelton and R. Woods (forthcoming 2006). Infant mortality: a continuing social problem? Ashgate.

Garrett, E. and Davies, R. (2003). "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, 53-74.

Gutmann, M.P. (1977). "Reconstituting Wandre. An approach to semi-automatic family reconstitution." Annales de Démographie Historique: 315-41.

Higgs, E. (2005). Making sense of the census revisited : census records for England and Wales 1801-1901 : a handbook for historical researchers. London, Institute of Historical Research.

History and Computing (1992), Special issue on record linkage, 4.1.

History and Computing (1994), Special issue on record linkage II, 6.3.

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

Holmes, R. (2006). 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.

Huntington P, Nicholas D, Jamali HR. (Forthcoming 2007). Employing log metrics to evaluate search behaviour and success: case study the BBC search engine. *Aslib Proceedings*, 59(1).

Impedovo, S. (1993). Fundamentals in Handwriting Recognition. NATO Advanced Study Workshop on Fundamentals in Handwriting Recognition, Château de Bonas, France, Springer-Verlag.

Inman, P. (2002). "Genealogy". The Guardian, Thursday September 26, 2002, <http://www.guardian.co.uk/internetnews/story/0,,798781,00.html>. Accessed 3rd November 2006.

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

International Organization for Standardisation (ISO) (2005). "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=39612&ICS1=35&ICS2=40&ICS3=> Accessed November 13th 2006.

Katz, M. and Tiller, J. (1972). "Record-linkage for everyman: A semi-automated process." Historical Methods Newsletter 5:144-150.

Kelly, G. A. (1955). The Psychology of Personal Constructs. New York, W.W. Norton and Company Inc.

Kingston Local History Project (2006), "The Kingston Local History Project" <http://localhistory.kingston.ac.uk/contributePages/klhp.html>, Accessed 3rd November 2006.

Lloyd, D. Webber, R. Longley, P. (2004). "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>. Accessed 9th November 2006.

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

Marcus, S., Ed. (1988). Automating Knowledge Acquisition for Expert Systems. Lancaster, Kluwer Academic Press.

McGraw, K. L. and Harbison-Briggs, K. (1989). Knowledge Acquisition: Principles and Guidelines. London, Prentice-Hall International Editions.

McGraw, K. L. and Westphal, C. R. (Eds). (1990). Readings in Knowledge Acquisition, Current Practices and Trends. London, Ellis Horwood Limited.

Mills, D., Pearce, C. Davies, R., Bird, J., and Lee, C. (1989) 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.

Morik, K., S. Wrobel, J-U. Kietz, and Emde, W. (1993). Knowledge Acquisition and Machine Learning: Theory, Methods and Applications. London, Academic Press.

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

Nitsch, D., Morton, S. DeStavola, B. L., Clark, H. and Leon, D. A., (2006) "How good is probabilistic record linkage to reconstruct reproductive histories? Results from the Aberdeen children of the 1950s study" BMC Medical Research Methodology, 6:15

Nicholas D., Huntington P., Jamali H.R. and, Tenopir, C. (2006) "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): 119-126.

Perkyns, A. (1991). "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, Leopard's Head Press.)

Perkyns, A. (1993). "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, Leopard's Head Press.)

The Programme de Recherche en Démographie Historique (PRDH). (2000). The 1852 and 1881 Historical Censuses of Canada, 1881 Cleaning Manual, Phase 1. Available at <http://www.prdh.umontreal.ca/1881/en/1881manual.html> Accessed 14/11/06.

Reichgelt, H. and Shadbolt, N. (1992). "ProtoKEW: A Knowledge-Based System for Knowledge Acquisition." In Sleeman, D. and N. Bernsen (Eds). Artificial Intelligence. Volume 6. Hove, Lawrence Erlbaum.

Reid, A., Davies, R. and Garrett, E. (2006) "Nineteenth century Scottish demography from linked censuses and civil registers: a 'sets of related individuals' approach", History and Computing, 14.1.

Robey, D. (2006) "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.

Sauleau, E. A., Paumier, J-P, and Buemi, A. (2005) "Medical record linkage in health information systems by approximate string matching and clustering". BMC Medical Informatics and Decision Making, 5:32.

Schürer, K. and Woollard, M. (2002). "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> Accessed 9th November 2006.

Shadbolt, N. and Burton, M. A. (1990). "Knowledge Elicitation Techniques - Some Experimental Results." In McGraw, K., L. and C. R. Westphal, (Eds). (1990).

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

Stallings, W. (forthcoming, 2007) Computer Security: Principles and Practice, Prentice Hall.

Terras, M. (2005). "Reading the Readers: Modelling Processes Used by Humanities Experts", Literary and Linguistic Computing, Volume 20. 41 - 59.

Terras, M. (2006). Image to Interpretation: An Intelligent System to Aid Historians in Reading the Vindolanda Texts. Oxford Studies in Ancient Documents, Oxford University Press. Oxford.

Tilley, P. (2003a). "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, P. (2003b). "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, P. M. (1972). "Sources of inaccuracy in the 1851 and 1861 censuses". In Wrigley (ed) (1972), Nineteenth-century society. Essays in the use of quantitative methods for the study of social data. Cambridge, Cambridge University Press. 82-133.

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

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

Warwick, C., Terras, M., Huntington, P., and Pappa, N. (Forthcoming 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." Submitted.

Waterman, D. A. (1986). A Guide to Expert Systems. Reading, Massachusetts, Addison-Wesley.

White, S. (2000). Enhancing Knowledge Acquisition with Constraint Technology. DPhil Thesis. Department of Computer Science. Aberdeen, University of Aberdeen.

Winchester, I. (1970). "The linkage of historical records by man and computer: Techniques and problems." Journal of Interdisciplinary History, 1: 107-24

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

Woollard, M. (1997). " 'Shooting the nets': a note on the reliability of the 1881 census enumerator's books". Local Population Studies, 59: 54-57.

Appendix A

The following appendices include the information given out to attendees on workshop days, for future reference.

A.1 Workshop 1: All Hands Workshop

A.1.1 Programme

AHRC e-Science Workshop in the Arts and Humanities

ReACH: Researching e-Science Analysis of Census Holdings

All Hands Workshop, Wednesday 14th June 2006

Foster Court 243, University College London

INTRODUCTION

e-Science allows 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 datasets, and little understanding of or access to e-Science technologies. The ReACH workshop series will 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.

This workshop series, based in the School of Library, Archive, and Information Studies at UCL, will feature input from

- 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)
- 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
- UCL Research Computing, the UK's Centre for Excellence in networked computing
- Experts in history, archives, genealogy, computing science, and humanities computing.

The All Hands Workshop, the first in the series, will aim to ascertain how feasible, and indeed, useful utilizing e-Science technologies to analyse historical census data will be. The workshop aims to bring

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 will explain the technological issues, and explore the historical techniques which may be useful for undertaking research of historical census material in this manner.

Results from this workshop will contribute to the discussions to be held at the forthcoming Technical and Managerial workshops. Undertaking e-science analysis of historical census records may be technically possible – but will it be useful to academic researchers?

PROGRAMME

| | |
|-------|--|
| 9.30 | <i>Coffee</i> |
| 10.00 | <i>Welcome and Introduction – aims of the day</i> Introducing ReACH Melissa Terras, SLAIS, University College London |
| 10.45 | Research Computing at UCL – An Overview Clare Gryce, Research Computing, University College London |
| 11.15 | Putting the Census Online: The National Archives' Perspective Ruth Selman, Knowledge and Academic Services Department, The National Archives |
| 12.00 | <i>Lunch</i> |
| 13.00 | Grid Enabling Population Datasets - the ConvertGrid and GEMS projects Keith Cole, Census Data Unit, National Dataset Services Group, MIMAS, The University of Manchester |
| 13.30 | Linking Nineteenth-Century Scottish Records: Procedures and Practicalities Ros Davies, Eilidh Garrett and Alice Reid, Cambridge Group for the History of Population and Social Structure |
| 14.00 | <i>Coffee</i> |
| 14.30 | Ancestry Census Records - Background, Structure, and Format Mike Wolfgramm, Vice President of Development, MyFamily |
| 15.00 | Discussion Session: e-Science Analysis of Historical Census Records: Feasible or Useful? In this discussion session all participants will be asked their opinions of the potential research project. Will undertaking this analysis result in any new information of worth for the academic community? What potential pitfalls are there in undertaking this research? What type of results can be generated from the available datasets? Is undertaking a research project in this area worthy of the time and expense it will take to set it up? |
| 16.00 | <i>Summary and Conclusion</i> |
| 16.30 | <i>Close</i> |

SYNOPSIS OF PAPERS

Introducing ReACH

Melissa Terras, SLAIS, University College London

There has been very little application of e-Science technologies in the arts and humanities to date. This paper will present an overview of the type of technologies described by the term “e-Science”, and how these may be appropriated to facilitate novel research in the Humanities. The background to the ReACH workshop series will be sketched out, and prime participants introduced. By presenting an overview of e-Science technologies, and the type of applications which may be useful with regards to historical census day, this paper aims to provide the background knowledge necessary for all participants to usefully contribute to further discussions throughout the day.

Research Computing at UCL – An Overview

Clare Gryce, Research Computing, University College London

UCL's Research Computing infrastructure supports a diverse profile of research from across all faculties. A variety of architectures, each suitable for handling different types of computational problem, offer services to meet the needs of UCL's growing community of Research Computing users.

UCL researchers are also active in national e-Science projects spanning the physical and engineering sciences, biomedical, life and social sciences. UCL has received over £18 million in funding as part of the UK e-Science Programme and the European Union's Information Society Technologies Programme.

This brief introduction to UCL's Research Computing activity will highlight some current e-Science projects at UCL, introduce the Research Computing community and provide an overview of the services available to all researchers - including those in the arts and humanities.

Putting the Census Online: The National Archives' Perspective

Ruth Selman, Knowledge and Academic Services Department, The National Archives

Ruth Selman is Knowledge and Information Manager in the Research, Knowledge and Academic Services Department at The National Archives (TNA). She has a degree in history from the University of Cambridge and has pursued her interest in history by studying her own and others' family history for the last 20 years. She has had a number of roles at TNA since joining in 1999, including acting as Transcription Manager for the 1901 Census project in partnership with QinetiQ.

Ruth will explain the scope of the census holdings of TNA and will touch on the Victorian statistical analysis of these as reflected in Parliamentary Papers. She will give an overview of the digitised census material currently available online and will focus specifically on the metadata available through Ancestry's service. Finally she will comment on the comprehensiveness and accuracy of both the original and the transcribed data.

Grid Enabling Population Datasets - the ConvertGrid and GEMS projects

Keith Cole, Census Data Unit, National Dataset Services Group, MIMAS, The University of Manchester

ESRC, JISC and other organizations, such as ONS, have invested substantially in the establishment of a distributed social science data infrastructure providing access to a range of key quantitative datasets, spanning many disciplines and research themes. These data and information resources play an increasingly important role in providing the evidence base for social science research. However, the increasing availability of data has resulted in a proliferation of different interfaces designed to deliver data to the user's desktop in a variety of different formats. Although adequate for ad hoc data extractions, the present situation is not conducive to systematic cross-analysis of multiple datasets; simplifying complex work flows; facilitating collaborative working; automating complex analyses that must be rerun periodically nor even to the retrieval and analysis of very large datasets. One major obstacle to the development of e-research in the social sciences is the absence of Grid enabled access to many of the key quantitative datasets, such as the aggregate census statistics. With reference to the ESRC funded ConvertGrid project and the JISC fund GEMS project this presentation will highlight how the establishment of a social science data Grid can automate very complex workflows and thus make it easier for researchers to investigate more complex research questions.

Ancestry Census Records: Background, Technology, Structures

Michael L. Mansfield, Director of Search and Content Engineering, Ancestry.com

Introduction

Ancestry's recent release of the 1841 Census of England, Wales, Isle of Man, and Channel Islands completes a multi-year project to digitize, key, and publish on-line the available genealogically useful decennial national censuses from 1841 to 1901. In association with the National Archives of England and Wales, this project has yielded a collection consisting of 164 million records of individuals enumerated by these seven decennial national censuses. Additionally, these records are linked to digital images of 6.75 million census schedules. This presentation will focus primarily on the keyed indexes and the techniques, technologies, search practices, and data structures employed at Ancestry to create, refine, and search these census collections.

Background & Motivation

From its roots in the 1970's as a print publisher of books and periodicals in the genealogy domain, Ancestry has focused on servicing the family history and genealogy market place. In 1995 the legacy print publication business was restructured to shift its focus to operating a subscription-based genealogical database and search service on the World-Wide Web. In early 1996, Ancestry.com was launched. From the beginning the customer demand for genealogical records has been insatiable and among the highest priority records are national census records. Ancestry.com experienced success in large part to its continued focus on imaging, indexing, and publication of the federal census records of the United States. In 2002, when the opportunity arose to work in cooperation with the National Archives of England and Wales to digitize and index the national census records of England and Wales the company readily embarked on the four year project.

Ancestry Census Records: Background, Technology, Structures (cont)

Digitization & Site Publication Overview

The methods and technologies used to scan the census microfilm, key the census schedules, audit and normalize the keyed index, and produce the formats consumed by Ancestry's search engine will be discussed. The census records are made available by The National Archives on microfilm. Digitization of the microfilm is done using high-speed microfilm scanners. The resultant digital images are then converted and compressed into JPEG 2000 images and shipped to keying partners. A complex keying application initially developed by Ancestry and continually enhanced by our keying partners serves as the primary tool by which the selected fields of the census schedules are extracted. The keyed contents of completed batches are returned to Ancestry where they under go a paleographic accuracy audit. Batches which fail the audit are returned to the partner for re-key. Once a complete collection has been keyed and all batches pass audit, it then goes though a normalization process. After normalization the index and images are prepared for online publication which includes such tasks as schema mapping, browse table generation, metadata and bibliographic creation, UI development, search system configuration and indexing, quality assurance, and site integration.

As the data flows through this content production pipeline it transitions through several data structures. An overview of these data structures and the processes which occur on them will be reviewed as part of the content production discussion. Additionally, a summary statistical report on the Censuses of England, Wales, Isle of Man, and the Channel Islands will be presented for the seven available decennial censuses from 1841 to 1901.

Searching Census Records

The presentation will culminate on the primary topic -- issues specific to searching, linking, stitching, and cross-referencing census records. The limitations of the extant records and accuracy of the keyed indexes will be presented along with well established domain best-practices for searching the census records using the keyed indexes. Specifically the caveats associated with names, dates, places, and relationships in census data will be presented and illustrated. These caveats will provide a foundation for a discussion of a genealogical relevancy model developed at Ancestry. This relevancy model is based on proximity measures for names, dates, places, and relationships. These search technologies constitute first generation approaches to mitigate the problems inherit in any genealogical dataset and to increase precision and recall. Additionally computational inferences, persona aggregation, data augmentation, and record stitching will be covered as advanced automated search techniques which are in various stages of research, development, and deployment at Ancestry.

Conclusion

The decennial national census indexes from 1841 to 1901 form a unique and large corpus. While this collection has obvious genealogical value to family history and genealogy researchers, additional research on the corpus is warranted. A myriad of studies and research topics out-side of the genealogy domain appear to be practical and necessary to evolve our understanding in many areas in demography, sociology, social-statistics, and social-economics to name a few of the many possible research fields.

A.1.2 Detail of Census Field available

Ancestry provided details of the fields available in their datasets across the different census years. Note that occupation and address has not been digitised in most cases due to the extra costs associated with digitising these fields (which are some of the most useful information for historians). On the following page, a guide to the available datasets are given, which was provided for workshop participants for analysis prior to the workshops. This informed discussion on the day.

| <i>Ancestry.co.uk</i> | | Available fields in datasets: England | | | | <i>Ancestry.co.uk</i> | |
|-----------------------------|-----------------------------|---------------------------------------|-----------------------------|------------------------------------|-----------------------------|-----------------------------|--|
| 1841 | 1851 | 1861 | 1871 | 1881 | 1891 | 1901 | |
| Name: | Name: | Name: | Name: | Name: | Name: | Name: | |
| Age: | Age: | Age: | Age: | Age in 1881: | Age: | Age: | |
| Estimated birth year: | Estimated birth year: | Estimated birth year: | Estimated birth year: | Estimated birth year: | Estimated birth year: | Estimated birth year: | |
| Relation: | Relation: | Relation: | Relation: | Relationship to head-of-household: | Relation: | Relation: | |
| Household: | Household: | Household: | Household: | Household: | Household: | Household: | |
| Gender: | Gender: | Gender: | Gender: | Gender: | Gender: | Gender: | |
| Where born: | Where born: | Where born: | Where born: | Where born: | Where Born: | Where born: | |
| Civil parish: | Civil parish: | Civil parish: | Civil Parish or Township: | Civil parish: | Civil parish: | Civil parish: | |
| Ecclesiastical parish: | Ecclesiastical parish: | Ecclesiastical parish: | Ecclesiastical parish: | Ecclesiastical parish: | Ecclesiastical parish: | Ecclesiastical parish: | |
| County/Island: | County/Island: | County/Island: | Town: | County/Island: | County/Island: | County/Island: | |
| Source information: | Source information: | Source information: | Source information: | Source information: | Source information: | Source information: | |
| Registration district: | Registration district: | Registration district: | Registration district: | Registration district: | Registration district: | Registration district: | |
| Sub-registration district: | Sub-registration district: | Sub-registration district: | Sub-registration district: | Sub-registration district: | Sub registration district: | Sub-registration district: | |
| ED, institution, or vessel: | ED, institution, or vessel: | ED, institution, or vessel: | ED, institution, or vessel: | ED, institution, or vessel: | ED, institution, or vessel: | ED, institution, or vessel: | |
| Folio: | Folio: | Folio: | Folio: | Folio: | Folio: | Folio: | |
| Page: | Page: | Page: | Page: | Page: | Page: | Page: | |
| Household schedule number: | Household schedule number: | Household schedule number: | Household schedule number: | Household schedule number: | GSU Number: | Household schedule number: | |
| GSU Number: | GSU Number: | GSU Number: | GSU Number: | GSU Number: | GSU Number: | GSU Number: | |
| | | | | | | 2.95GB | |

A.1.3 Attendees

The following persons attended the workshop on the 14th June, contributing to both papers and discussion.

| | |
|---|--|
| Josh Hanna | Managing Director and Vice President, Ancestry Europe |
| Mike Mansfield | Director of Content Engineering and Search, MyFamily Inc |
| Ruth Selman | Knowledge and Information Manager, The National Archives |
| Dan Jones | Licensing Manager, TNA (Manager of relationship with Ancestry) |
| Chris Owens | Head of Access Development Services, TNA Manager, UCL Research Computing and e-Science Centre of Excellence, |
| Dr Claire Gryce, Professor David Nicholas | Department of Computer Science, UCL Chair of Library and Information Studies, UCL: SLAIS |
| Dr Melissa Terras | Lecturer in Electronic Communication, UCL:SLAIS |
| Dr Claire Warwick | Lecturer in Electronic Communication and Publishing, UCL:SLAIS |
| Dr Andrew MacFarlane | Lecturer, Department of Information Science, City University Director of the Census Data Unit, Deputy Director of National Dataset |
| Dr Keith Cole | Services Group, MIMAS, The University of Manchester |
| Professor Rob Procter | Research Director of the National Centre for e-Social Science |
| Dr Edward Higgs | Reader, Department of History, University of Essex. Co-ordinator, Centre for Scholarly Editing and Document Studies |
| Edward Vanhoutte | (KANTL), Ghent |
| Professor Kevin Schürer | Director of the Economic and Social Data Service (ESDS) and the UK Data Archive (UKDA), Department of History, University of Essex. |
| Richard Holmes | MA Research Student, UCL |
| Dr Tobias Blanke | Arts and Humanities e-Science Support Centre |
| Dr Jeremy Yates | UCL research computing, Lecturer in Physics and Astronomy |
| Dr Ros Davies | Cambridge Group for the History of Population and Social Structure |
| Dr Alice Reid | Cambridge Group for the History of Population and Social Structure |
| Dr Eilidh Garrett | Cambridge Group for the History of Population and Social Structure |
| Dr Eccy de Jonge | UCL SLAIS |
| Dr Kevin Ashley | Head of Digital Archives, University of London Computer Centre |
| Dr Matthew Dovey | Technical Manager, Oxford E-science Centre, University of Oxford |
| Duncan MacNiven | Registrar General for Scotland |
| Geoffrey Yeo | Lecturer in Archives and Record Management, UCL SLAIS |

Pablo Mateos

Department of Geography / CASA, University College London

A.2 Workshop 2: Technical Workshop

A.2.1 Programme

AHRC e-Science Workshop in the Arts and Humanities

ReACH: Researching e-Science Analysis of Census Holdings

Technical Workshop, Thursday 15th June 2006

South Wing Council Room, University College London

INTRODUCTION

The second workshop in the ReACH series will build on the conclusions from the All Hands Meeting (14th June 2006, UCL). Participants are 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 will provide the technical information necessary for inclusion in any further funding bid regarding implementation of a full scale project.

This workshop meeting will aim 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 (access to existing kit)
- Software required
- Software development issues
- Security issues
- Any other technical issues which may arise

PROGRAMME

| | |
|-------|---|
| 9.30 | <i>Coffee</i> |
| 10.00 | <i>Welcome and Introduction – aims of the day</i> Melissa Terras, SLAIS, University College London |
| 10.15 | Round table discussion session regarding technical issues |
| 12.30 | Summary and Close |
| 13.00 | <i>Lunch</i> |

A.2.2 Attendees

| | |
|----------------------|---|
| Josh Hanna | Managing Director and Vice President, Ancestry Europe |
| Mike Mansfield | Director of Content Engineering and Search, MyFamily Inc |
| Ruth Selman | Knowledge and Information Manager, The National Archives |
| Dr Melissa Terras | Lecturer in Electronic Communication, UCL:SLAIS |
| Dr Andrew MacFarlane | Lecturer, Department of Information Science, City University |
| Dr Tobias Blanke | Arts and Humanities e-Science Support Centre |
| Dr Jeremy Yates | UCL research computing, Lecturer in Physics and Astronomy |
| Dan Jones | Licensing Manager, TNA |
| Chris Owens | Head of Access Development Services, TNA (ICT Specialist) Manager, UCL Research Computing and e-Science Centre of Excellence |
| Dr Claire Gryce | Department of Computer Science, UCL |

A.3 Workshop 3: Managerial Workshop

A.3.1 Programme

AHRC e-Science Workshop in the Arts and Humanities

ReACH: Researching e-Science Analysis of Census Holdings

Management Workshop, Tuesday 25th July 2006

Taviton Room 431, University College London

AGENDA

| | |
|------|--|
| 2.30 | <i>Welcome and Introductions</i> |
| 2.45 | ReACH so far: an overview (Melissa Terras) |
| 3.00 | Discussion session |
| 4.30 | Close - Wine |

Overview

The ReACH workshop series aims 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.

The managerial workshop is the third (and final) workshop to be undertaken as part of this research series. Previous workshops have addressed the academic aims, and technical infrastructure necessary to develop a grid based project which utilises historical census data. The aim of this workshop is to ascertain the managerial and legal issues which will 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 will be discussed will include

- Licensing requirements from Ancestry
- Security of data
- Ownership of research outcomes

- Management structure of a full scale project
- Financial structure of a full scale project
- Paths to dissemination and publicity
- Any other topics which participants suggest are relevant

The workshop will mostly be a discussion session, featuring input from Ancestry.co.uk, The National Archives, UCL SLAIS, UCL Research Computing, and Historians.

A.3.2 Attendees

| | |
|--------------------------|---|
| Josh Hanna | Managing Director and Vice President, Ancestry Europe |
| Dan Jones | Licensing Manager, TNA |
| Chris Owens | Head of Access Development Services, TNA |
| Dr Claire Gryce | Manager, UCL Research Computing and e-Science Centre of Excellence, Department of Computer Science, UCL |
| Professor David Nicholas | Chair of Library and Information Studies, UCL: SLAIS |
| Dr Melissa Terras | Lecturer in Electronic Communication, UCL:SLAIS |
| Dr Jeremy Yates | UCL research computing, Lecturer in Physics and Astronomy |
| Dr Eddy de Jonge | UCL SLAIS |
| Dr Tobias Blanke | Arts and Humanities e-Science Support Centre Director of the Economic and Social Data Service (ESDS) and the |
| Professor Kevin Schürer | UK Data Archive (UKDA), Department of History, University of Essex. (Expert in census analysis) |

A.4 Steering Group

A.4.1 Programme

AHRC e-Science Workshop in the Arts and Humanities

ReACH: Researching e-Science Analysis of Census Holdings

Steering Committee, Wednesday 19th July 2006

Taviton Room 431, University College London

AGENDA

| | |
|-------|---|
| 12.30 | <i>Lunch</i> |
| 12.45 | <i>Welcome and Introductions</i> Report on Activities so far Discussion regarding future activities – final workshop Discussion on reporting and conclusions Discussion regarding aspects for future funding Budget |
| 2.00 | Close |

Report

The ReACH workshop series aims 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.

This workshop series, based in the School of Library, Archive, and Information Studies at UCL, has featured input from

- 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)
- 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
- UCL Research Computing, the UK's Centre for Excellence in networked computing
- Experts in history, archives, genealogy, computing science, and humanities computing.

The team working on the ReACH workshop series comprises of Dr Melissa Terras, with administration support from Dr Eccy de Jonge. The aim of the workshop series is to ascertain how and why e-Science analysis of census holdings may be undertaken, taking from this pilot project aspects which may affect anyone wishing to undertake e-Science grid type analysis with complex humanities data. It is also hoped that the workshop series will result in a concrete proposal to put forward to future funding rounds. Three workshops are planned: the Alls Hands workshop, the Technical Workshop, and the Managerial workshop, which will aim to investigate the different issues in undertaking such analysis.

The All Hands Workshop, the first in the series, was undertaken on 14th June 2006. It aimed to ascertain how feasible, and indeed, useful utilizing e-Science technologies to analyse historical census data will

be. 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. The workshop was well attended, by 25 individuals from across the historical, archival, and computing science domains. Discussion was lively, and was recorded. In depth notes were taken throughout. Feedback following the workshop has been universally positive.

The Technical workshop, a much smaller and more focused affair attended by 10 delegates, took place on Thursday 15th June, and featured input from Ancestry, TNA, and computing scientists and physicists from the UCL Research Community, aiming to ascertain throughput and workflow of a potential project, estimate technical requirement and staffing requirements, and come up with approximate costings for a project. The meeting, again, was lively, and by the end of the discussion a proposal was sketched for a 2 year project: this requires more work but can be discussed at the meeting.

The final workshop, the Managerial workshop, will take place on July 25th, and will involve representatives from the institutions involved to discuss issues such as security, copyright, ownership of project outcomes, etc. This will be attended by 10 delegates.

The website has not yet gone live due to institutional delays - however, a password and url have finally been obtained (www.ucl.ac.uk/reach/) and this will be implemented over the next week!

The main costs to the project have been travel expenses and sustenance. Given that the all hands and technical workshops were run back to back, we managed to save on travel costs for the delegate from Ancestry.com coming from the USA. The grant for this project is approximately £10,000, and we are currently under budget.

After the final workshop, the main aspects of the research remaining will be to report on the project, and to ensure that the project results are disseminated. The project wraps up towards the end of August, and will report back to the AHRC in written form. Papers emanating from this conference will be submitted to relevant conference in History, e-Science, and Digital Humanities in the autumn.

A.4.2 Steering Group

| | |
|-------------------------|---|
| Dr Melissa Terras | Lecturer in Electronic Communication, UCL:SLAIS Arts and Humanities e-Science Support Centre |
| Dr Tobias Blanke | Lecturer in Archives and Record Management, UCL SLAIS Manager of Arts and Humanities, The London E- Science. Centre |
| Geoffrey Yeo | Centre for Computing in the Humanities, King's College London Communications Manager, Arts and Humanities Data Service |
| Dolores Iorizzo | Knowledge and Information Manager, The National Archives Manager, AHRC ICT Methods Network |
| Martyn Jessop | Licensing Manager, TNA Head of AHDS History, Project Director of the Online Historical Population Reports Project |
| Alastair Dunning | Director of the Economic and Social Data Service (ESDS) and the UK Data Archive (UKDA), Department of History, University of Essex. |
| Ruth Selman | |
| Lorna Hughes | |
| Dan Jones | |
| Dr Matthew Woollard, | |
| Professor Kevin Schürer | |