International Workshop Population Reconstruction
Amsterdam, The Netherlands

Programme and abstracts

Day 1, Wednesday 19 February 2014 chair: Gerrit Bloothooft

09.00 - 09.30 **Registration and coffee**

09.30 - 10.00 **Opening**
Kees Mandemakers, *IISH, Amsterdam*

10.00 - 11.00 **Keynote Peter Christen,** *Research School of Computer Science, Australian National University, Canberra*

Advanced record linkage methods and privacy aspects for population reconstruction

Recent times have seen an increased interest into techniques that allow the linking of records across databases. The main challenges of record linkage are (1) scalability to the increasingly large databases common today; (2) accurate and efficient classification of compared records into matches and non-matches in the presence of variations and errors in the data; and (3) privacy issues that occur when the linking of records is based on sensitive personal information about individuals. The first challenge has been addressed by the development of scalable indexing techniques, the second through advanced classification techniques that either employ machine learning or graph based methods, and the third challenge is investigated by research into privacy-preserving record linkage. In this paper, we describe these major challenges of record linkage in the context of population reconstruction, outline recent developments of advanced record linkage methods, and provide directions for future research.

11.00 - 11.20 ***Coffee break***

Linking theory

11.20 - 12.05 Graham Kirby, Conrad de Kerckhove, Ilia Shumailov, Jamie Carson, Alan Dearle, *School of Computer Science, University of St Andrews, Scotland*; Chris Dibben, Lee Williamson, *Longitudinal Studies Centre Scotland, Universities of St Andrews and Edinburgh, Scotland*

Comparing Relational and Graph Databases for Pedigree Data Sets

Increasingly large family pedigree data sets are being constructed from routine civil and religious registration data in various parts of the world. These are then being used in health, social and genetic research in a variety of different ways. Often the type of questions that are being asked involve complex queries, such as the degree of relatedness between multiple sets of individuals, and involve traversing through the data typically multiple times. There is therefore an important issue of efficiency of querying. In this paper we evaluate the suitability of two classes of database, relational (MariaDB) and graph (Neo4j), for storing and querying pedigree datasets representing millions of individuals. We report results of measurements of scalability, query performance, and the ease of query expression, using synthetic datasets.

12.05 - 12.50 Corry Gellatly, *Research Institute for History and Culture, Utrecht University, The Netherlands*

Reconstructing historical populations from genealogical data: an overview of methods used for aggregation data from GEDCOM files

The GEDCOM file format is by far the most widely used means of exchanging genealogical data and extensive collections of these files are available online. There is a huge potential benefit for historians and other academics who are able to make use of the data contained in available GEDCOM files, as these effectively represent hundreds of thousands of hours of crowd-sourced work and a considerable source of knowledge about individual families. This paper details a number of methods that are being used to clean and aggregate such genealogical data; this includes a series of steps for screening out substantially flawed files, as well as for cleaning date and place information. A group-linking method is described for identifying duplicates / linkages within a genealogical database based on comparison of family structures. This is tested alongside conventional methods (i.e. comparison of name and birth date) and an estimation of the power of the differing methods is provided. It is proposed that use of the group-linking method provides advantages over conventional methods, because this provides a way of increasing the size and timespan of datasets that may be extracted from a genealogical database with confidence that they do not contain duplicates. The method will be further improved by incorporating probabilistic record linkage techniques, which take into account the frequencies of values in the linkage arrays.

12.50 - 14.00 ***Lunch***

14.00 - 14.45 Julia Efremova[1], Bijan Ranjbar-Sahraei[2], Frans A. Oliehoek[2], Toon Calders[1/3], Karl Tuyls[2/4] , [1]*Eindhoven University of Technology, The Netherlands;* [2]*Maastricht University, The Netherlands;* [3]*Université Libre de Bruxelles, Belgium;* [4]*University of Liverpool, United Kingdom*

A Baseline Method for Genealogical Entity Resolution

In this paper we study the application of entity resolution (ER) techniques on a real-world multi-source genealogical dataset. Our goal is to identify all persons involved in various notary acts and link them to their birth, marriage and death certificates. In order to evaluate the performance of a baseline approach based on existing techniques, an interactive interface is developed for getting feedback from human experts in the field of genealogy. We perform an empirical evaluation in terms of precision, recall and F-score. We show that the baseline approach is not sufficient for our purposes and discuss future improvements.

Group linking

14.45 -15.30 Trygve Andersen, Gunnar Thorvaldsen, *Norwegian Historical Data Centre, University of Tromsø*

Record linkage in the Historical Population Registry for Norway

A national population registry will become a unique historical source for the last two centuries and may be used in many different research projects. The potential inherent in the rich Nordic source material will be released once the nominative records are linked together in order to describe persons, families and places longitudinally. This requires the development of new linkage techniques combining both automatic and manual methods, consisting of a composite of several established techniques combined with new methods that increase the overview. It will require a large concerted effort from a large number of people to achieve these goals, but in return this gives us an infrastructure with great utility for many different research projects. We are cooperating with several international projects in the field of computerizing nominative sources, and look forward to future exchanges of ideas, especially on how to better use combinations of the church registers in population research. From an international perspective, there are no comparable open and national historical population registers built by linking multiple source types.

15.30 - 15.50 ***Tea break***

15.50 -16.10 Marijn Schraagen, *LIACS Leiden, The Netherlands*

Historical record linkage using event sequence consistency

Databases with personal information, such as hospital records or historical archives, generally contain multiple records involving the same individual person or group of individuals. The process of identifying sets of records involving the same people (or entities in general) is known as record linkage. In this paper the consistency of a sequence of records is used as linkage criterion, in order to reduce dependence on pairwise string similarity measures. The approach is applied to a database of Dutch historical civil certificates containing 4 million birth certificates, 3 million marriage certificates and 7.5 million death certificates. Evaluation is performed using a manually created benchmark (22,000 people) of family reconstructions.

16.10 - 17.00 **General discussion on linking theory**

17.00 **Drinks**

Day 2, Thursday 20 February 2014

chair: Peter Christen

09.30 - 10.30 **Keynote Arno Knobbe,** *Leiden Institute of Advanced Computer Science, Leiden, The Netherlands*

Reconstructing Medieval Social Networks from English and Latin Charters

The ChartEx project has been developing new ways of analyzing historical documents in an integrated fashion, and reconstructing medieval social networks based on this analysis. Specifically, the project's aim was to develop tools to deal with medieval charters: records of legal transactions of property of all kinds: houses, workshops, fields and meadows. The project was started by an international consortium under the Digging into Data program, bringing together historians and experts in Natural Language Processing, Data Mining and Human Computer Interaction. Partners from the following countries were involved: Canada, United Kingdom, United States, and the Netherlands. The role of the NLP experts was to automatically annotate the large collections of charters for further processing, based on example annotations of a sample of charters produced by the historians. The data miners then proceeded to link charters based on the actors and sites mentioned, and thus reconstruct some of the social network embedded in the collections. This will be the main focus of this paper.

10.30 - 11.00 ***Coffee break***

Data cleaning and standardization

11.00 - 11.45 Nanna Floor Clausen, *Danish Data Archive, Danmark*

Danish Demographic Database – principles and methods for cleaning and standardization of data

In this article is presented the work that has been done and is still being done in cleaning and standardizing seven completely transcribed Danish censuses. The transcriptions are done by volunteers and with the use of different applications. The information in the censuses are not standardized and the transcriptions have further added variation to the original data making cleaning and standardization a tiresome though interesting task.

11.45 - 12.30 Ivo Zandhuis, *Ivo Zandhuis Research & Consultancy*; Menno den Engelse, *Islands of Meaning*; Edward Mac Gillavry, *Webmapper, The Netherlands*

Dutch historical toponyms in the Semantic Web

In early 2013, the Dutch website gemeentegeschiedenis.nl was launched. The website presents a uniquely identifiable web page (a so called "uri") for every municipality in the Netherlands since 1812. Each of these web pages provides internal relations to official spelling alternatives of their names, to toponyms of settlements within these municipalities, between former and current municipalities, and presents maps of all the geographical changes to the official boundaries of these municipalities. In this paper we present an evaluation of the information presented on the website and a description of the services needed to create an even more useful tool. This, in order to demonstrate the usefulness of a centralised information system for historical toponyms.

12.30 - 13.15 Graham Kirby, Jamie Carson, Fraser Dunlop, Alan Dearle, *School of Computer Science, University of St Andrews, Scotland*; Chris Dibben, Lee Williamson, *Longitudinal Studies Centre Scotland, Universities of St Andrews and Edinburgh, Scotland*; Eilidh Garrett, Alice Reid, *Department of Geography, University of Cambridge, UK*

Automatic Methods for Coding Historical Occupation Descriptions to Standard Classifications

The increasing availability of digitized registration records presents a significant opportunity for research in many fields including those of human geography, genealogy and medicine. Re-examining original records allows researchers to study relationships between factors such as occupation, cause of death, illness, and geographic region. This can be facilitated by coding these factors to standard classifications. This paper describes work to develop a method for automatically coding the occupations from 29 million Scottish birth, death and marriage records, containing around 50 million occupation descriptions, to standard classifications. A range of approaches using text processing and supervised machine learning is evaluated, achieving accuracy of 92.3 ± 0.2% on a smaller test set. The paper speculates on further development that may be needed for classification of the full data set.

13.15 - 14.00 ***Lunch***

14.00 - 14.45 Gerrit Bloothooft, *Utrecht University, The Netherlands*; Marijn Schraagen, *LIACS Leiden, The Netherlands*

Learning name variants from true person resolution

Name variants which differ more than a few characters can seriously hamper person resolution. A method is described by which variants of first names and surnames can be learned to a large extent automatically from records that contain more information than needed for a true link decision. Limited manual intervention (active learning) is unavoidable, however, to differentiate errors from variants in the original and the digitized data. The method is demonstrated on the basis of an analysis of 14.8 million records from the Dutch vital registration (from 1811 onwards).

Short papers on resources

14.45 - 15.05 Alexander Buczynski, Vedran Klaužer, *Croatian Institute of History, Department of Early Modern History, Croatia*

Muster rolls of the Croatian Military Frontier as sources of historical demography

This paper deals with 18th and 19th century muster rolls from the Military Frontier in Croatia as extraordinary valuable sources of historical demography. The authors will discuss the possibilities, advantages and shortcomings of these military lists for demographic research, and present some preliminary solutions for data entry based on sources like these into a longitudinal database.

15.05 - 15.25 XingChen C.C. Lin, *Department of History, TamKang University, Taiwan*

Challenge and prospect of combining Taiwanese historical information of social stratification and occupation database (THISCO) into HISCO

In this initial working paper, the author describes the procedure and problems of incorporating the occupational information from household registers, which titles are described in the book *LinShi Taiwan HuKou DiaoCha ZhiYe MingZiHui,* into the Historical international classification of occupations (HISCO), to create a Taiwanese Historical Database with Information of Social Class and Occupations (THISCO). Currently, the historical occupational information has been completely digitalized, and contains 6,817 records in Mandarin. After digitization, THISCO moves toward an international linkage stage, and faces some problems that need to be solved. The paper will report several differences between Taiwanese occupational titles and HISCO, and presents a possible solution.

15.25 - 15.45 ***Tea break***

15.55 - 16.05 Antero Ferreira, Carlota Santos, *CITCEM-GHP, Universidade do Minho, Guimarães, Portugal*

National Genealogical Repository - Developing a Central Database

This paper presents the National Genealogical Repository project which intends, in a first phase, merge all the existing genealogical databases on the Portuguese population and, in a second phase, aims to broaden its focus nation-wide. We will also discuss some methodological questions related to this process of integration.

16.05 – 16.25 Karin Hofmeester, Rombert Stapel, Richard Zijdeman*, IISH, Amsterdam, The Netherlands*

Global Collaboratory on the History of Labour Relations, 1500-2000

Today's rise of precarious work, as highlighted in scientific and political debate, is part of a much longer change in the division of labour within societies. The Global Collaboratory on the History of Labour Relations main goal is to provide insight in the distribution of populations across labour relations (systematically including women's and child labour) around the world in five historical cross-sections: 1500, 1650, 1800, 1900, and 2000. In our presentation we will briefly explain the setup of the project: a Collaboratory in which a large number of international scholars work together to create a dataset with data on labour relations. We will give an overview of the type of data we gather, amongst others population data for all regions included, the sources we use and our methods to attribute labour relations and to visualize our first results with the use of tree maps.

16.25 – 17.00 **General discussion on cleaning and standardization, and resources**

19.00 - 23.00 **Workshop dinner**
Brasserie Beems, Rokin 74, Amsterdam

Day 3, Friday 21 February 2014

Chair: Kees Mandemakers

09.30 - 10.30 **Keynote Kris Inwood**
Luiza Antonie, Kris Inwood, J. Andrew Ross, *University of Guelph, Canada*

Dancing with dirty data: Problems in the extraction of life-course evidence from historical censuses

This paper builds on a recent use of SVM classification to create linked sets of Canadian 1871 and 1881 census records (Antonie et al., 2013). This method generates lifecourse information for large numbers of individuals. The new records are of considerable research value although characteristics of the original data create challenges. Feature taken from the census have limited granularity and therefore many people share identical detail. Reporting is imprecise. In spite of these challenges the false positive error rate of the linked records is only 3%. However there is a higher incidence of error among apparent migrants when the true rate of migration is small. The linked data are broadly representative of the population with some under-representation of illiterates, young adults (especially unmarried women), older people (especially men), and married people of French origin. Research use of the new longitudinal data must take into account these characteristics.

10.30 - 11.00 ***Coffee break***

Life courses

11.00 - 11.45 Janet McCalman, Rebecca Kippen, Sandra Silcot, *Centre for Health & Society, University of Melbourne, Australia*; Leonard Smith, *Demographic and Social Research Institute, Australian National University*

Building a Life Course Dataset from Australian Convict Records; Founders & Survivors: Australian Life Courses in Historical Context, 1803-1920

Founders & Survivors is a multi-university and public collaborative project that is building a transnational and inter-generational dataset of life courses generated from the UNESCO recognized convict records of Tasmania. This paper outlines the technical history of the project: mass digitization and archiving online of over 100,000 images; manual scholarly transcription; TEI standard XML data library based on automated and manual record matching and linkage; crowdsourcing using Google Docs to manage over fifty-four online volunteer genealogists; reconstitution of amalgamated life courses and record linkage; development of customized genealogical database for population and family analysis (Yggrasil); export to statistical programs (SPSS, Stata). Manual linkage and scholarly verification remained essential for the collation of prosopographical data and manual coding, derived from historical analysis was necessary for statistical analysis.

11.45 - 12.30 Francisco V. Goula, *Max Planck Institute for Demographic Research, Rostock, Germany*; Joan-Pau Jordà, Joana M. Pujadas-Mora, *Centre for Demographic Studies (CED), Autonomous University of Barcelona, Spain*

Reconstructing lifespans using historical marriage records of Catalonia from the 16th and 17th centuries

This paper presents a methodology and data process to reconstruct the lifespan of individuals using historical marriage records. The reconstructed lifespans are being used in an ongoing research project to estimate adult life expectancy in Catalonia in the 16th and 17th centuries. The data set to be used is the Barcelona Historical Marriage Database (BHMD) which collects marriage license records for a continuous period of almost five hundred years (1451-1905). The document discusses the main characteristics of the original source that has enabled the building of the BHMD, the harmonization process of the database, and the nominal record linkage process through which we link marriage records of grooms and brides with their respective parents in order to reconstruct genealogies and individual lifespans.

12.30 - 13.30 ***Lunch***

Short papers on linking strategy

13.30 - 13.50 John Bass, Sandra Silcot, Len Smith, *University of Melbourne, Australia*

Founders and Survivors linkage strategy

The prosopography database supporting the Founders and Survivors project comprises a relational genealogy database integrated with an XML source database. Individual life histories are compiled dynamically from diverse sources, linked by a combination of machine matching and human judgment, and managed by an independent link management module. It is planned to enhance the machine linkage using ontology-based semantic web technology.

13.50 - 14.10 Vlad Popovici, '*Babeş-Bolyai' University, Centre for Population Studies, Cluj-Napoca, Romania*

Preliminary Steps for the Reconstruction of Transylvania's Population in the mid-18th–early 20th Centuries

This short paper aims at highlighting the main difficulties related to the process of building the first large digital database of the population from Transylvania (mid-18th to early 20th centuries), as well as presenting the solutions envisioned in order to ensure the success of the project in the long run.

14.10 - 14.30 Catalina Torres, *Programme de recherche en démographie historique (PRDH), Université de Montréal, Canada*

Issues regarding the use of the Canadian census sample of 1852 for data linkage

This paper discusses some issues that should be considered when using the Canadian census sample of 1852 in data linkage projects.

14.30 - 14.50 Alice Reid, University of Cambridge, UK; Eilidh Garrett, University of St Andrews, Scotland

Introducing 'movers' into community reconstructions: linking civil registers of vital events to local and national census data: a Scottish experiment.

This paper describes our experience of tracing individuals and family groups observed in the 1861-1881 civil registers of Skye and the 1861 and 1871 censuses of the island to their places of residence elsewhere in Scotland on the date of the 1881 census.

14.50 - 15.10 ***Tea break***

15.10 - 15.40 **General discussion on life courses and linking strategy**

15.40 - 16.10 **Conclusions and closure**