Project Summary

This proposal seeks continued funding for the North Atlantic Population Project (NAPP). This project is a unique collaboration of seven leading data producers who have leveraged resources to create an extraordinary cross-national historical database. Over the past four years, the collaborating partners have worked to clean, edit, code, harmonize, and disseminate almost 90 million records describing basic demographic characteristics of the populations of five countries. These data include the entire population of Britain and Canada in 1881, Iceland in 1870, 1880, and 1901, Norway in 1865 and 1900, and the United States in 1880. These are the only complete-count national microdata available for scholarly research, and they represent an extraordinary resource for the study of small areas and population subgroups. This fundamental social science infrastructure will stimulate broad-based comparative investigations of economic development and demographic change.

To exploit the research potential of these data, the second phase of NAPP will (1) expand the chronological dimension of the database by incorporating data from additional census years for each country; (2) link individuals between censuses to permit longitudinal analysis; and (3) improve the web-based tools for disseminating data and documentation.

Intellectual merit. As it stands, the NAPP database is a remarkable resource, permitting for the first time the study of small areas, dispersed subgroups, and minority populations in a critical period of economic and demographic transition. The potential, however, is even greater. The NAPP database presently provides a cross-sectional snapshot of the characteristics of five countries in the late nineteenth century. In each country, however, additional machine-readable data are available for other census years. Most of these are national samples of censuses, as opposed to the complete-count data currently in the NAPP database. The new project will add 23 additional censuses to the database to allow the study of social and economic change in the nineteenth and early twentieth centuries. The investigators will then link individuals in the samples to the complete-count NAPP database, allowing individual-level longitudinal analysis of social mobility, family transitions, migration, and a host of additional topics. Finally, the project will adapt and implement new tools for data dissemination and on-line analysis currently being developed for another NSF-funded project.

The international teams are located at the leading centers of population history infrastructure in each country, and represent an extraordinary pool of talent and expertise. Each institution has obtained local resources to construct separate national data series. Funding from NSF allows NAPP to harness these resources to create a combined database that is far more powerful than the sum of its parts. The NAPP project is a rare success story of international collaboration. Producing a single coherent database with staff and funding scattered across seven institutions on two continents requires continuous intensive communication and negotiation. This is hard work, and without resources specifically dedicated to nurturing the collaboration, it would be impossible.

Broader impact. The availability of multiple cross-sections for the population of the North Atlantic world in the nineteenth and early twentieth centuries will open up vast new terrain in the fields of history, economics, demography, and sociology. In addition, linked samples hold the promise of finally resolving some of the longest-running debates in social and economic history. Scholars will be able to gauge the extent of social and geographic mobility, the interrelationship of geographic and economic movement, and trends and differentials in social mobility far more reliably than heretofore.

NAPP has already become an important component of social science infrastructure. At this writing, over 270 projects by scholars at over 130 institutions are underway. Reducing the technical barriers to using the data by adopting new web-based data access and analysis tools will broaden the audience considerably. The availability of on-line data analysis tools will make important contributions to teaching in the social sciences, helping to bring the excitement of discovery into the classroom.

TABLE OF CONTENTS

For font size and page formatting specifications, see GPG section II.C.

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.

Objectives

This proposal seeks continued funding for the North Atlantic Population Project (NAPP). The project is a unique collaboration of seven leading data producers who have leveraged resources to create an extraordinary cross-national historical database. Over the past four years, the collaborating partners have cleaned, edited, coded, harmonized, and disseminated 90 million records describing basic demographic characteristics of the populations of five countries. These data include the entire population of Britain and Canada in 1881, Iceland in 1870 and 1901, Norway in 1865 and 1900, and the United States in 1880. These are the only complete-count national microdata available for scholarly research, and they represent an extraordinary resource for the study of small areas and population subgroups. This fundamental social science infrastructure is already stimulating broad-based comparative investigations of economic development and demographic change.

To exploit the research potential of these data, we now propose (1) expanding the chronological dimension of the database by incorporating data from additional census years for each country; (2) linking individuals between censuses to permit longitudinal analysis; and (3) improving our web-based tools for disseminating data and documentation.

The NSF funding for NAPP (SES-0111707) did not provide the funds to collect, clean, or code the data; rather, we requested funds to coordinate existing national projects. A dozen funding agencies in five countries contributed resources to develop individual national census databases. The concept of NAPP was that a small investment in collaboration among these countries would leverage these assets and ensure that the full potential of the data is realized. Because of the NAPP, almost all the data are now freely available to scholars in harmonized format through a web-based data extraction system.

As it stands, the NAPP database is a remarkable resource, permitting for the first time the study of small and dispersed population subgroups in a critical period of economic and demographic transition. The potential, however, is even greater. The NAPP database presently provides a cross-sectional snapshot of the characteristics of five countries in the late nineteenth century. In each country, however, additional machine-readable data are available for other census years. Most of these are national samples of censuses, as opposed to the complete-count data currently in the NAPP database. We propose to add 23 additional censuses to the database to allow the study of social and economic change in the nineteenth and early twentieth centuries. We will then link individuals in the samples to the complete-count NAPP data, allowing for individual-level longitudinal analysis of social mobility, family transitions, migration, and a host of additional topics. Finally, we will adapt and implement new tools for data dissemination and online analysis that are currently being developed for another NSF-funded project.

Participants in each country have obtained national resources to clean, code, and process the additional datasets and to construct linked samples that allow longitudinal analysis.[1] Like the original NAPP initiative, the purpose of this proposal is to fund costs associated with coordinating our efforts to create a single, compatible cross-national database. The close coordination of database construction across seven research centers requires significant effort that cannot be funded exclusively through the separate national projects. We therefore propose to extend our successful collaboration, and dramatically expand the power of the NAPP database at marginal cost.

Introduction to revised proposal

This is a resubmission of a proposal submitted in August 2004. The proposal received strong reviews from both the Sociology and the MMS review panels; of eight reviewers, four rated the proposal as "excellent," three rated it "very good," and one rated it "good." Both panels recommended funding at the medium priority level,

[1] In addition to the funding described in the original NAPP proposal, in the past two years participants have been awarded additional grants to work on the project. These include Kevin Schurer, "Victorian Panel Study" (Economic and Social Research Council, RES-500-25-5001, $183,000); Lisa Dillon, "Historical Demography Research Infrastructure" (Canadian Foundation for Innovation, New Opportunities Fund 7549, $383,850); Lisa Dillon, "Les transitions aux familles québécoises" (Fonds de recherche sur la société et la culture, Établisement de nouveaux professeurs-chercheurs 90225, $40,500); Lisa Dillon, "Canadian Households Across Time and Space" (Social Sciences and Humanities Research Council of Canada, Standard Research Grant 58917, $71,448); Chad Gaffield, "Canadian Century Research Infrastructure" (Canadian Foundation for Innovation, $5.2 million); Steven Ruggles, "1880 United States Population Database: Continuation" (NIH HD39327, $2.6 million); Gunnar Thorvaldsen, "Name Standardization in Nineteenth-Century Censuses" (Norwegian Research Council). In addition, Statistics Iceland is allocating internal resources to this effort.

but the agency requested a resubmission to provide additional methodological detail requested by several of the reviewers. The thoughtful reviewer comments have resulted in a significantly improved proposal. To make room for the additional technical material, we have reduced the detail on background and significance.

Three reviewers requested additional information about our record-linkage procedures. One reviewer noted correctly that linkage accuracy is as important as representativeness, and we have altered our text to clarify our desire to maximize both; it is the linkage rate, rather than accuracy, that we are prepared to sacrifice. The reviewers requested further information on sources for the control totals for weighting, the estimated sizes of the linked samples, the problems of name cleaning across languages, and several other details. We have addressed these specific questions in the text of the proposal, but record linkage is complex and a full description is impossible within the space constraints of the grant proposal. Instead, we provide references to the key methods developed by statistical agencies and computer scientists, and use the bulk of our space to highlight the specific ways our approach differs from current standard practice. As we note, additional details are available in the two papers we have written on our proposed methods, Ferrie (2003) and Ruggles (2003b), available at http://www.nappdata.org/imag.shtml. Even there, however, we do not fully document the specific implementation of these methods for each country, for the simple reason that the implementation is not yet finalized; refinement of the linking procedures represents a significant component of the research project.

Two reviewers questioned whether it is worth including Iceland in the project. One of these referred to the very *small size* of the Icelandic datasets, and suggested that the datasets are so small that they would not be suitable for statistical analysis. We estimate that the Icelandic data—comprised of the complete enumerations from six census years—will include approximately 373,000 cases, which is a sufficient number to allow in-depth statistical analysis. A second reviewer expressed concern about the *quality* of the Icelandic data, particularly with respect to occupations. In fact, however, the Icelandic censuses may be the highest quality data incorporated in NAPP, and the occupational data is particularly good, providing detail on secondary occupations as well as primary occupations in later census years. Iceland is a key element of the project; it has more complete-count censuses than any other country in the world, and this collection offers greater potential for longitudinal analysis than any other component of the NAPP database.

One reviewer questioned the selection of the NAPP countries and did not see the utility of adding additional census years. The choice of countries for the original project was straightforward: NAPP includes all the countries for which there were completely digitized individual-level censuses for the late nineteenth century. The great majority of countries do not have surviving enumeration manuscripts for their nineteenth-century censuses, thus prohibiting this kind of research project. Moreover, the close economic integration of the NAPP countries and the high migration flows among them also provide a compelling rationale for studying the region. We perhaps did not make clear that the chronological dimension of the database is essential for the study of each of the topics highlighted in the expected significance section of the proposal, such as industrialization, the fertility transition, and the transformation of family and household composition.

Finally, the MMS panel summary requested additional detail on the sample designs that were used for the 19 samples we propose to add to the database. The 1851 British sample is clustered (by enumeration district), but all the rest are high-precision systematic designs with implicit geographic stratification. As part of the documentation for the project, we will prepare estimates of design effects for the samples. Since we do not use place of residence as a linking variable, the clustered design of the 1851 census does not improve the ease of record linkage. Space constraints do not permit full documentation of the various sample designs, but we have provided a brief summary and additional references to the full documentation.

Results of prior NSF support

The NSF project most closely related to the present proposal is the "North Atlantic Population Project." (SES-0111707, 08/01/2001-07/30/2005, $491,506), but the project builds on 15 years of NSF-funded experience integrating large census microdata samples. Starting in 1991, Ruggles developed the Integrated Public Use Microdata Series (IPUMS) to harmonize, document, and disseminate all U.S. census microdata (SES-9118299, SBR-9422805, SBR-9617820). In 1999, Ruggles began extending the IPUMS paradigm to international censuses of the past four decades (SBR-9908380). The international IPUMS project will be continued for another five years through a major grant from the NSF initiative on Human and Social Dynamics (SES-0433654). The expanded IPUMS-International will include recent census data for at least three of the five NAPP countries (Britain, Canada, and the United States), and we are actively pursuing agreements to disseminate data for the other two (Iceland and Norway). These recent data will greatly enhance the value of

NAPP for the study of economic and demographic change over the very long run. In addition, the new IPUMS project will support development of web-based data dissemination tools that we can easily adapt to the NAPP.

The scale of the NAPP database is very large; with almost 90 million observations, NAPP is currently larger than either IPUMS-USA or IPUMS-International. This is clearly "Big Science," but it does not have a big science price tag; the cost to NSF is less than a tenth as much as the IPUMS projects. This partly reflects the major contributions of funding agencies in each country to the project. In large measure, however, it results from our ability to leverage the past investment of NSF and NIH in infrastructure originally developed for the IPUMS projects.

The five NAPP countries were selected because they all possess completely digitized individual-level censuses for the late nineteenth century. The Church of Latter-Day Saints (LDS), in collaboration with local genealogical societies, laboriously digitized three of these censuses—for Britain, Canada and the United States—to provide a resource for genealogical research. That massive project involved some 12 million hours of work by thousands of volunteers and professionals, and resulted in a verified transcription of the census information on the population of those countries in 1880 or 1881. Over the past two decades, Norwegian researchers have invested more than half a million hours in digitizing historical population records as a source for social science research. In Iceland, nineteenth-century censuses were transcribed as part of an effort to construct genealogies for genetic research.

The result of all these labors was a transcription of the characteristics of 90 million persons who resided on the North Atlantic rim in the late nineteenth century. The census in each case provides information on age, sex, marital status, family relationships, occupation and birthplace, and allows the construction of a full complement of variables describing household composition, fertility, and neighborhood and community characteristics. In their raw form, however, these data were of minimal use for social science research. There are literally millions of occupational titles, birthplaces, family relationships and geographic localities transcribed in four different languages. Before any of these data could be exploited for social science, researchers had to numerically code, classify, and document each variable.

In Britain, Canada, Norway, and the United States, researchers pieced together funding from numerous sponsors to support the painstaking tasks of data cleaning and coding. In Britain, the funders include the Economic and Social Research Council, the Leverhulme Trust, and the Essex University Research Promotion Fund; in Canada, the Social Sciences and Humanities Research Council, les Fonds de recherche sur la société et la culture, the Canadian Foundation for Innovation, the Harold Crabtree Foundation, the Church of Jesus Christ of Latter-Day Saints, and the University of Ottawa Research Partnerships Programme; in Norway, the Norwegian Research Council, the Norwegian National Archives, and the Faculty of Social Sciences of the University of Tromsø; and in the United States, the National Science Foundation and the National Institutes of Health.

The researchers knew that if they coordinated their efforts, the datasets could be pooled, allowing cross-national analyses of the North Atlantic population. Initial discussions about the potential for creating an integrated complete-count census database occurred in Ottawa in April 1999, at a meeting of the International Microdata Access Group (IMAG). In June and October 2000 participants from each country met in Minneapolis to define the goals of the project and develop a detailed plan of work. The participants agreed that we should not simply create compatible datasets, but rather should develop a single fully integrated database with common coding systems, constructed variables, documentation, and dissemination systems.

Although each collaborator obtained funding to process their own national censuses, there were no resources to support the intensive collaboration that was needed to ensure that the data would be compatible across countries. We therefore proposed NAPP, which provided funding to cover costs associated with coordinating international harmonization of the data. Most of the collaboration was carried out via the Internet, but NAPP also provided funds for a series of workshops at which we hammered out solutions to the most complicated issues.

The scale of our task was daunting. To give just one example, the collaborating partners coded over two million different occupational titles in four languages into a common classification scheme. To maximize cross-national consistency in coding, thousands of occupational titles were independently classified by researchers from multiple countries, and discrepancies in coding were resolved in conference. Were it not for NAPP, each

country would have coded occupations into a different national classification, and cross-national comparison would have been impossible.

The project has been extraordinarily successful. We released preliminary versions of the U.S. data in August 2003 through our web-based data access tool, the Canadian data went online in December 2003, and the British and Norwegian data in November 2004. We anticipate that the Icelandic data—which required substantially more work than we expected—will be incorporated into the database before the end of the project in 2005. The database is distributed through http://www.nappdata.org, and the final version—incorporating all variables described in the original proposal—will be released on schedule in July 2005.

NAPP has already become an important component of social science infrastructure. At this writing, over 270 research projects by scholars at over 130 institutions are underway, even though the data have only been readily available for 18 months. This is comparable to the number of users of IPUMS in 1995, at a similar stage of development (today, there are over 14,000 IPUMS users). Despite the brief period the data have been available, some important research has already appeared, including seven published or forthcoming articles, two working papers, and 17 conference presentations (http://www.nappdata.org/publications.shtml). Among these exciting papers is a study by Sacerdote (forthcoming) demonstrating that that it took roughly two generations for the descendants of slaves to catch up to the descendants of free black men and women with respect to literacy, occupational attainment, and schooling. Potentially even more important, a preliminary analysis by Ferrie and Long (2004) suggests that social mobility was exceptionally high in nineteenth-century America compared to both nineteenth-century England and to late-twentieth-century America; if confirmed, this study will overturn much of the literature on the history of social mobility in America.

The majority of the researchers are faculty, graduate students, or postdoctoral researchers at major universities. Despite the massive size of the database, however, undergraduates have also been using the data for senior papers and theses. Most researchers using NAPP are based in the five countries whose data is available, but we also have users from Australia, Austria, Denmark, Germany, Japan, the Netherlands, Spain, Sweden and Switzerland. Research topics include the living arrangements of the aged, female labor-force participation, seasonality of births, European migration to North America, migration between Canada and the United States, divorce and separation, mortality forecasting, history of entrepreneurship among blacks after Reconstruction, comparison of the black population in the Great Lakes regions of the United States and Canada, community studies, and the history of specific occupational and industrial groups (e.g. barbers, railroad porters, pottery manufacturers, hospitality trades).

We have published three articles about the methods employed to construct the NAPP database: Evan Roberts, Steven Ruggles, Lisa Dillon, Ólöf Garðarsdóttir, Jan Oldervoll, Gunnar Thorvaldsen, and Matthew Woollard (2003) "The North Atlantic Population Project: An Overview." *Historical Methods* 36 (2): 80-88; Evan Roberts, Matthew Woollard, Chad Ronnander, Lisa Dillon, and Gunnar Thorvaldsen (2003) "Occupational Classification in the North Atlantic Population Project." *Historical Methods* 36 (1): 89-96; and Woollard, Matthew (2004) "The Classification of Multiple Occupational Titles in the 1881 Census of England and Wales." *Local Population Studies* 72: 34–49. In addition, we have presented 12 conference papers about technical aspects of the project (http://www.nappdata.org/publications.shtml).

Expected significance

The first objective of this proposal is to expand the chronological scope of the database. Although the present NAPP database provides a resource of unprecedented power for fine-grained comparative analysis of nineteenth-century populations, it is of limited use for understanding change over time. For Britain, Canada, and the United States, NAPP presently includes a single census year; for Norway, it includes two census years, and for Iceland, it includes three. Raw machine-readable data from 23 additional censuses are available, and we propose to incorporate these files into the database.

The addition of data for multiple census years in each country will multiply the potential applications of the NAPP database. The availability of multiple cross-sections for the population of the North Atlantic world in the nineteenth and early twentieth centuries will open up vast new terrain in the fields of history, economics, demography, and sociology. This is a critical period in the study of fertility decline, urbanization, international migration, household composition, and occupational structure. A full discussion of the specific topics that could be addressed with the addition of a chronological dimension to the NAPP database would require many pages. The paragraphs that follow sketch only a few of the most obvious research applications of the new data.

Industrialization. The first Industrial Revolution may have begun in Lancashire, but by the late nineteenth century the entire North Atlantic world was involved in manufacturing, the production of raw materials, or both. The North Atlantic database will allow unprecedented opportunities to explore economic structures within and between each nation during this critical transitional period. For the first time, we will have consistently coded occupational data available for multiple censuses in five countries, and it will be available at the individual level. Four of the five countries, for example, had mechanized textile manufacturing, and the census provides sufficient occupational detail to analyze the changing organization of this key industry.

Fertility transition. In the second half of the nineteenth century, the population of each of the North Atlantic countries was just beginning deliberate fertility limitation.[2] The North Atlantic database will allow study of differential fertility patterns in this critical period of demographic transition, to assess the importance of such factors as occupational class, ethnicity, region, literacy, local economy, size of locality, and family structures. Study of this elemental shift in population structure has the potential to enhance our understanding of ongoing demographic change in the contemporary developing world.

Past comparative analyses of the European fertility transition have relied on aggregate vital statistics (Coale and Watkins 1986). This approach has two major disadvantages. First, aggregate vital statistics do not allow direct measures of child spacing or stopping behavior; only the *level* of fertility can be considered. Second, the aggregate approach does not allow control of individual-level socioeconomic characteristics. The new database will allow analysis of fertility trends and differentials through own-child methods (Cho, Retherford and Choe 1986). Thus, the database will allow new and more subtle comparative analyses of the first demographic transition across five nations.

Household and family composition. For more than a century, political theorists, sociologists and historians have been debating the relationship between industrialization and the family. In the 1970s, a series of British, Canadian and American studies argued that the harsh economic conditions of early industrial capitalism strengthened the interdependence of family members and led to a high frequency of complex households (Anderson 1972; Hareven 1978, 1982; Katz 1975; Foster 1974; Modell 1978). Each of these analyses focused on a single industrializing community, so they were unable to test the proposed association between industrial development and family or household composition. By allowing study of household composition across time and space simultaneously, the expanded NAPP database will permit more powerful comparative analyses of changing household and family composition than was previously possible.

The North Atlantic database will include a common set of constructed variables to aid in the analysis of family and household composition and will thus allow consistent comparisons across all five countries. These variables will allow investigators to assess the impact of local context on family systems through multilevel analysis, and for the first time permit analysis of the effects of individual-level factors, local economic conditions, regional inheritance systems, and national characteristics on changes in the family.

Linked samples. The availability of cross-sections from multiple census years in each country will make it feasible to study social and economic change, but the database will still consist of a series of cross-sectional snapshots. To address this limitation, we propose to link the censuses across time, and in some instances across countries, to provide multiple observations for the same individual.

The NAPP research teams in Britain, Canada, Norway, and the United States have each obtained funding to link their respective complete-count censuses to samples of other nineteenth and early twentieth-century censuses, and Statistics Iceland has also agreed to devote resources to this endeavor. Using new record-linkage technology, we will construct samples linking individuals in the complete-count census year to the available census samples, creating multiple pairs of linked datasets. In the United States, for example, we plan a series of linked samples for 1850-1880, 1860-1880, 1870-1880, 1880-1900, 1880-1910, 1880-1920, and 1880-1930. Each of the paired samples will be independent, but taken together they will provide a rich longitudinal source for the nineteenth and early twentieth centuries.

[2] Until recently, demographers thought that American fertility decline began much earlier than in most of Northern Europe. Hacker (1999, 2000a, 2003), however, has shown using census-based own-child and back projection methods that deliberate fertility limitation in the United States began considerably later than was previously thought. Hacker places the fertility transition in America after the Civil War, which is the same period that it occurred in each of the other countries in the North Atlantic database.

We propose to develop automated linking algorithms that can be used to construct linked samples for each participating country. Because we will be using virtually the same algorithms in each country, this approach will maximize cross-national comparability, but differences in the source data prevent us from guaranteeing that the characteristics of the linked samples will be absolutely identical. Linked samples will greatly enhance the value of the previous investments in the construction of the NAPP database. If we do not closely coordinate these efforts, however, the resulting datasets will not be comparable.

Many of the most important studies using the NAPP data—such as the Sacerdote (forthcoming) and Ferrie and Long (2004) papers noted above—are based on record linkage between the sample data and the complete-count data. It does not make sense, however, for each scholar to develop and implement *ad hoc*, idiosyncratic and non-replicable linking strategies whenever longitudinal panel data are needed. By taking advantage of the most recent developments in record-linkage technology, we can produce high-quality standardized samples that will reduce the cost of research and improve the reliability of results.

The linked samples hold the promise of finally resolving some of the longest-running debates in social and economic history. Scholars will be able to gauge the extent of social and geographic mobility, the interrelationship of geographic and economic movement, and trends and differentials in social mobility far more reliably than heretofore. Past studies of these topics were ultimately inconclusive because of their exclusion of migrants and their small sample size. In addition, the linked samples will allow investigation of family formation and dissolution. For example, they will allow us to settle an important debate about the formation of multigenerational households in the nineteenth century (Ruggles 1994, 2003a), an issue with important implications for the study of intergenerational relations and the twentieth-century transformation of the living arrangements of the aged. Because the linked samples will be comparable across countries, it will be feasible for the first time to assess systematically national differences in economic opportunity, migration, and family transitions.

In addition to linking censuses within countries, we plan to experiment with record linkage across countries. Where possible, we will follow individual migrants across the Atlantic. The late nineteenth century saw international population movements on an unprecedented scale. The massive North Atlantic migration profoundly shaped both the receiving and contributing countries. The great majority of emigrants from Norway, Iceland and Britain went to Canada and the United States, and the influx transformed North American society. Many of these newcomers remained only a few years before returning to their homelands, often bringing home money and always bringing new ideas and experiences (Runblom and Norman 1976; Nugent 1992; Gjerde 1992; Thorvaldsen 1997). The internationally-linked census samples will open a new window on immigration history by allowing us to assess changes in the characteristics of migrants at the individual level.

Improved data access. The third goal of this project is to improve the tools for accessing NAPP data and documentation and to develop a system that will allow online data analysis. Web-based data extraction software is essential because of the scale of the NAPP data; the entire database would require about 100 CD-ROM disks. We are presently using a data extraction tool descended from software we developed in 1996 for the original IPUMS project. It allows users to select variables, subset populations, and merge data from multiple countries and census years. Data are delivered in ASCII format with setup files for SAS, SPSS, or Stata. Despite its success, this system nevertheless has significant limitations: it is comparatively slow, it does not allow users to save and edit previous extracts, it does not permit complex case selections, and it does not construct new variables on the fly. Developing a new data access tool specifically for NAPP that addresses these limitations would be prohibitively expensive. Fortunately, NSF is funding the IPUMS-International project to develop a new generation of data access software that will overcome these problems. Adapting the new IPUMS-International software to the NAPP will provide a cost-effective means to improve access to the database.

We also plan to implement basic online data analysis software for accessing the cross-sectional data. This initiative—which also builds on IPUMS-International software development—will allow users to create tabulations of NAPP data without downloading it or using a statistical package. Scholars can use the system for feasibility analyses; students, journalists, and others without access to statistical software will for the first time gain access to the resource. We anticipate that by reducing the technical barriers to data exploration, we will broaden our audience substantially. For example, historians are a natural constituency for NAPP data, but most lack the software and expertise presently needed to make use of it. We also expect that online data analysis tools will make important contributions to teaching in the social sciences, helping to bring the excitement of

discovery into the classroom. The chronological and geographic analysis made possible by expansion of the NAPP database makes it a suitable vehicle for introducing a quantitative dimension into secondary, undergraduate and graduate courses in a broad range of fields.

Methods and procedures

The project consists of three components. First, we will extend the chronological dimension of the NAPP database by adding data from additional censuses; second, we will create longitudinal samples by linking individuals across censuses; and third, we will upgrade the web-based data access tools for NAPP and add online data analysis capabilities. After a brief description of our source data, the following sections depict our general approach in broad strokes. We then discuss documentation, sustainability, work plan, and evaluation.

Source data. Table 1 lists the data files to be included in the North Atlantic database. The current NAPP database includes seven censuses: the three census transcriptions created by the LDS for Britain, Canada and the United States in 1880/1881, and five files covering the Icelandic and Norwegian censuses between 1865 and 1901. We now propose to add 23 additional files.

For Britain, we will include a 2 percent sample of the 1851 census of England and Wales originally compiled by Michael Anderson in the 1970s (Anderson, Stott, and Collins 1979). Anderson never completed the cleaning and coding of this important dataset, but Matthew Woollard and Kevin Schurer of the U.K. Data Archive have already invested substantial effort to bring it up to current standards. There is also a strong possibility that a machine-readable version of the 1901 British census may become available during the life of this project.

We will incorporate Canadian data from five different sources: (1) a new sample of the 1852 census being developed by Lisa Dillon at the Université de Montréal, (2) a sample of the 1871 census created by Gordon Darroch and Michael Ornstein of York University, (3) a sample of the 1891 census now under construction by Kris Inwood of the University of Guelph, and (4) a 1901 sample created under the direction of Eric Sager and Peter Baskerville at the University of Victoria; and (5) anonymized samples now being produced for the period 1911 through 1951 by the Canadian Century Research Infrastructure (CCRI) project under the direction of Chad Gaffield. The samples from 1911 onward have confidentiality restrictions and will be subject to disclosure review by Statistics Canada before they can be made publicly available. To preserve confidentiality, these samples will not be part of the linking project at this time. The samples for 1871 and 1901 will require considerable retrofitting, but the other samples are being developed by members of the NAPP consortium, and we expect that they can be added to the NAPP database with minimal modification.

We plan to add complete-count data from two of the earliest surviving individual-level censuses in the world: the 1703 census of Iceland and the 1801 census of Norway. Once coded, these will provide unprecedented opportunities to assess demographic and economic behavior before industrialization. We will also incorporate three additional complete-count Icelandic censuses between 1835 and 1880; taken together with the Icelandic data already in NAPP, this will yield five complete-count censuses for Iceland at frequent intervals between 1835 and 1900. This will be an extraordinary source for record linkage, especially since Icelandic women do not change their surnames upon marriage. In addition to the 1801 census, the Norwegian series will be augmented by data from the 1875 census, which has complete coverage for a third of the municipalities and provides a two-percent sample of the rest. Finally, for the United States we will convert all the IPUMS samples that are not anonymized into NAPP format, including high-density samples of 1900 and 1930 now under construction.

With respect to their structure, organization, and available information, the 31 censuses to be incorporated in the NAPP database are remarkably comparable. In each case, the censuses describe the characteristics of individuals grouped into households, and the interrelationships of individuals within households can be determined. All countries defined households as a group of people sharing a common place of residence.

There is a core set of variables common to virtually all datasets, including relationship of each individual to the household head, age, sex, marital status, occupation, and birthplace. In addition, there are many important variables available for subsets of censuses, such as year of immigration, religion, and automobile ownership. As in the case of the IPUMS projects, our goal is to lose no information. Space does not permit a full listing of the availability of the 141 NAPP variables, but they are documented at http://www.nappdata.org/subjects.html.

The geographic units identified vary from country to country, mainly because of differences in the political organization of each nation. In all countries, however, we can identify the location of all places with 5,000 or

more persons, and we estimate that we can identify approximately 25,000 places across the five countries. The common core variables will allow us to construct a variety of new variables describing community and neighborhood characteristics, household composition, socioeconomic status, and family interrelationships.

Table 1. Existing and proposed NAPP datasets

Census Year	Country	Sample Density	Number of Records (thousands) Household	Person
EXISTING NAPP FILES				
1881	Great Britain	1.00	6,188	29,866
1881	Canada	1.00	799	4,278
1870	Iceland	1.00	11	60
1880	Iceland	1.00	14	72
1901	Iceland	1.00	15	78
1865	Norway	1.00	387	1,702
1900	Norway	1.00	395	2,294
1880	United States	1.00	10,138	50,486
TOTAL EXISTING			17,933	88,764
FILES TO BE ADDED				
1851	Britain	0.02	83	398
1852	Canada	0.05	31	170
1871	Canada	0.01	13	62
1891	Canada	0.05	67	350
1901	Canada	0.05	51	265
1911	Canada	0.05	74	372
1921	Canada	0.04	74	362
1931	Canada	0.03	67	320
1941	Canada	0.03	77	355
1951	Canada	0.03	93	420
1703	Iceland	1.00	9	50
1835	Iceland	1.00	10	56
1845	Iceland	1.00	10	57
1801	Norway	1.00	164	879
1875	Norway*	0.02	135	639
1850	United States	0.01	37	198
1860	United States	0.01	66	354
1870	United States	0.01	80	428
1880	United States	0.10	1,014	5,049
1900	United States	0.06	1,248	5,220
1910	United States	0.01	311	1,271
1920	United States	0.01	257	1,037
1930	United States	0.06	1,670	6,160
TOTAL TO BE ADDED			5,665	24,544

* For 1875 Norway, the density is 100% in 149 municipalities and 2% in 347 municipalities.

Most of the censuses were taken on a *de jure* basis, under which individuals who were temporarily absent from home—such as migrant workers and travelers—were to be enumerated at their usual place of residence. The exception is the British census, taken under a *de facto* rule which specified that no one who was present on census night at a particular address could be left out of the tally, and that no person absent from home could be written in. In Norway and Iceland from 1870 onwards, persons were to be enumerated both at their usual place of residence *and* at the place they stayed on census day, and enumerators identified both temporary visitors and absent household members. The Norwegian and Icelandic censuses will therefore allow us to assess the implications of enumeration rules for comparative study of household composition and migration issues.

Data quality is good. The United States census is probably the weakest of the group in this respect. The most recent demographic analysis indicates that net underenumeration of the U.S. 1880 census was 6.4 percent (Hacker 2000b; King and Magnuson 1995). Although coverage was not complete, the overall response rate to the American census in the late nineteenth century compares favorably with modern survey data, such as the Current Population Survey. We lack comparable estimates for Britain, Canada, Iceland and Norway, but because of their smaller size, more homogeneous populations, lower geographic mobility and stronger central governments, census taking was considerably less challenging.

Nineteen of the 23 datasets we propose to add to NAPP are samples, and there is some variation in the sample designs.[3] In general, the samples are designed to be nationally representative, proportionally weighted, and high-precision. For example, in the United States in 1910, a sampling window of five lines was randomly placed on every fifth page of the 100-line enumeration forms. Any unit of 30 or fewer individuals that began within the window was included in the sample, and any individual residing in a unit of 31 or more who fell within the sampling window was also included. This sample maximizes precision by limiting the size of clusters, and since the forms are geographically sorted it capitalizes on implicit geographic stratification. At the same time, the threshold of 31 persons for individual-level sampling ensures that all regular households are captured as intact units. The other U.S. and Canadian samples are all similar, although there are slight variations owing to variation in the source materials and to the progressive refinement of procedures over time. The Norwegian sample of 1875 is slightly different. For approximately one-third of the population—northern Norway, cities, and a representative selection of rural municipalities—the dataset includes the entire population. For the other two-thirds of the population, the 1875 dataset includes a systematic sample of every 50th dwelling unit. Although we have not yet undertaken a formal analysis of design effects for the Canadian and Norwegian samples, we expect that sample precision for those countries will be similar to the U.S. design.

The British sample design for 1851 differs significantly from that of the other countries: it is a stratified cluster sample based on enumeration districts. Enumeration districts, which contain an average of approximately 150 households, were first divided into six categories based on the size of place, urban or rural character, and presence of institutions. Within each category, the sample includes all residents of every 50th enumeration district. For institutions, every 50th individual was included in the sample. Thus, implicit geographic stratification ensures equal representation of each part of the country, and explicit stratification ensures that each type of place is represented. Because of clustering by enumeration district, however, the British sample is less precise than the samples of the other countries. We plan to estimate design effects for the British 1851 data by constructing sample replicates using 1851 sampling rules drawn from the 1881 complete-count database. We are confident that because the sample size is very large, the 1851 data will offer adequate precision for most analytic purposes. Moreover, because our linking strategy ignores place of current residence (see below), the difference in sample design will not affect record linkage.

Harmonizing variables. Space does not permit a full description of our approach to harmonization of nineteenth-century international microdata. We have published the details of our methods elsewhere (Roberts et al. 2003a, Roberts et al. 2003b). The following paragraphs briefly summarize our general approach.

The format of each of the datasets we plan to add to the NAPP database varies widely. In virtually every case, however, we have access to alphabetic character strings (in English, French, Icelandic or Norwegian) that represent a transcription of the information collected from each individual. The individual records are grouped into residential units corresponding to the modern census concepts of household and group quarters. Each country has raised funds to classify these alphabetic strings into numerically coded categories. Some variables—age, sex and marital status—can be made comparable with little effort, but the complex variables require close collaboration to develop common coding standards. As in the case of the original NAPP datasets, occupational coding is the most challenging component of the project. The fine detail available in the occupational field is one of the reasons why the North Atlantic database has the potential to transform our understanding of historical social structure. At the same time, however, the complexity of occupational structure demands meticulous care to ensure consistency. NAPP uses a modified version of the Historical International Standard Classification of Occupations (HISCO) as our basic framework for occupational classification (Edvinsson and Karlsson 1998; van Leeuwen, Maas and Miles 2002; Roberts et al. 2003b). The

[3] A full description of the design of each sample is not possible within the space constraints of this proposal; full descriptions are available elsewhere. See Hall, McCaa and Thorvaldsen (2000) for complete references to the sample designs.

HISCO system in turn is a modification of the 1968 United Nations occupational classification system with extensions to accommodate historical occupations, developed by an international committee with representatives from Belgium, Canada, England, France, Germany, the Netherlands, Norway, Sweden and the United States.

To translate from character strings into numeric codes, we constructed data dictionaries that assign a numeric code to each alphabetic variation that occurs in the data. This work is difficult enough in the context of a single country; for a project of this scale, it requires a team of expert coders who work in close cooperation, sharing coding decisions continuously. The merged NAPP dictionaries are of unprecedented scale, since they include the alphabetic strings from all five countries. The occupation dictionary, for example, includes 2,605,301 entries.

Our investment in these dictionaries will greatly ease the incorporation of the new datasets into NAPP. Many of the alphabetic strings we encounter will already appear in the merged NAPP dictionaries, and virtually all of them will be similar to previously encountered values. As in the case of the original project, our collaboration on the expanded data dictionaries will exploit Internet-based tools to update the merged dictionary database, verify coding decisions, and discuss differences of interpretation.

In addition to harmonizing variable coding, we will construct a range of compatible derived variables for each new dataset we incorporate into NAPP. These include technical variables to aid in data management and analysis, such as record type, serial number, group quarters residence and household size; variables describing urban/rural residence, size of place, and metropolitan residence; variables to describe the composition of families and households; and variables to aid in the analysis of family interrelationships and own-child fertility analysis.

Record linkage. Historians have been linking individuals across censuses for decades, but the results have been problematic. In most cases, linked census studies were based on local populations because no complete census for a larger area has been available. These studies generally lose between 60 and 80 percent of the population each decade due to linkage failures (see for example Anderson 1972, Katz 1975, Knights 1991, Oldervoll 1982, Thernstrom 1964, Thorvaldsen 1995). Most linkage failure is attributable to the very high migration of the mid-nineteenth century. The availability of a high-quality census database including entire national populations for the late nineteenth century will allow far more sophisticated matching than has previously been possible. Using the NAPP database in combination with recent advances in record-matching technology, entire countries can be searched using such characteristics as age, sex, birthplace, birthplace of mother, and birthplace of father, as well as name. The complete-count census data will allow a higher rate of matches than previous linkage studies, and will provide much larger linked samples. Even more important, because the linked samples will be constructed from representative populations at both ends of the record linkage, selection bias on observed characteristics in the linked population will be readily detectable and largely correctable through weighting.

We propose to exploit new record-linkage and data-mining technology to create linked representative samples of individuals and family groups. The National Institutes of Health, the Economic and Social Research Council of the United Kingdom, the Social Sciences and Humanities Research Council of Canada and the Norwegian Research Council have each funded projects to link the complete-count databases of each country to samples of surrounding censuses. The NAPP will piggyback on these efforts and create compatible linked samples for all five countries. The linking project for the United States samples is already well underway, and this work constitutes a rigorous pilot study for the larger international project. Space constraints do not permit a full description of our record-linkage strategies; further details are available in Ruggles (2003b) and Ferrie (2003), available at http://www.nappdata.org/imag.shtml. The following paragraphs, however, provide an overview of our general approach.

General linkage strategies. Although social scientists have linked historical records for over 50 years, during the past 15 years technological developments have opened new opportunities to create more powerful linked historical datasets than were previously possible (Rosenthal 1997, Committee on Applied and Theoretical Statistics 1999). Our procedures will build on these innovations. Our goals, however, differ significantly from those of most recent researchers. The primary goal of virtually all the work on record linkage has been to maximize the number of valid links. A typical data-mining application, for example, would involve linking membership records to address lists to identify potential sales prospects. The goal of such an application is not to create a statistically valid representative sample, but simply to generate the largest possible number of

customers. The most important linking application for statistical agencies is the estimation of census undercount through the capture-recapture method, so they also aim for the largest possible number of reliable links.

We will not focus on maximizing the linkage rate. Instead, our procedures will be designed to maximize the *representativeness* of the linked cases and the *accuracy* of the links. This means that we must pay close attention to sources of selection bias, and ignore much of the information routinely used by other record-linkage procedures. Although it is impossible to eliminate the possibility of selection bias for unobserved characteristics, we can adopt procedures that greatly reduce the potential for bias compared with previous approaches.

The principal applications of the proposed linked samples will be the study of social mobility, migration, family change, and life-course transitions. We therefore must avoid using any information that could bias the sample with respect to those changes. For example, record-linkage algorithms ordinarily make use of place-of-residence as a linking variable. This greatly increases the potential for reliable links: if we identify an individual in a sample who partially matches the name and other characteristics of a person in the complete-countdatabase, our confidence that the two records refer to the same person would be improved if we knew that they both reside in the same locality. If we use place of current residence in the linking algorithm, however, we will inevitably bias the sample in favor of non-migrants. Likewise, if we use spouse's name in the algorithm we will bias the sample in favor of those who remain married, and if we use occupation we will favor cases with low social mobility.

We plan three categories of linked samples, each with a different universe: all males, females who do not marry in the census interval, and married couples. Although we are linking individuals or couples, we will also capture all characteristics of all coresident household members. Even though none of these groups is representative of the entire population, our goal is to make each category representative of its defined universe. The male individual sample will be general purpose, useful for studying economic and geographic mobility, transitions to adulthood, changes in family composition, and retirement. The female sample will be useful for studying many of the same topics, but since it will be composed of the subset of women who do not change their surname between censuses, it will be inappropriate for some topics. The married-couple samples will offer the greatest reliability, since it will allow us to link on characteristics of both husband and wife, and will be especially useful for topics relating to fertility, child mortality, and age of leaving home. Because it is restricted to the continuously married population, however, it will be less useful for population-wide generalizations about social and geographic mobility.

For each sample, we will start by identifying a subset of individuals in the sample datasets (e.g., Britain 1851 or Canada 1891). We will then search for these individuals in the censuses that have complete coverage (e.g., Canada or Britain 1881). For the United States, Britain, and Canada there are a total of 13 pairs of censuses that we can link. In the case of Norway—which has complete-count data for 1865 and 1900 and sample data for 1875—we will link individuals in the 1875 sample to both the 1865 and 1900 census years. Iceland has five complete-count censuses between 1835 and 1901, and we will endeavor to locate the same individuals across all five census years.

We will create three linked samples for each pair or series of census years, for a total of 45 samples. Half the samples use forward links (e.g., 1871 to 1881) and half rely on backward links (e.g., 1891 to 1881). Forward-linked samples are more challenging than the backward-linked ones because mortality and emigration substantially reduce the potential for links. Backward linkage is only complicated by immigration, but this problem is simplified by the availability of retrospective information in Canada and the United States—the principal receiving countries—about year of immigration for the foreign-born population. Together with age, this question will allow us to define a universe that approximates the population that was alive and resident in the country in 1880 or 1881. Although this universe will be imperfect because of errors in enumeration and transcription, it will allow more aggressive linking strategies by reducing uncertainty.

Our algorithm will rely exclusively on characteristics that should not change over time. At minimum, these variables are first name, last name (for men and unmarried women), birth year, sex, and place of birth. For some countries, we have additional variables, such as race or parental birthplaces. Genealogists and data miners make use of a far broader range of characteristics to confirm links and resolve ambiguities. We believe,

however, that knowledge of any additional characteristics would introduce biases that would severely damage the samples.

The chief problem posed by our approach is that this limited set of variables is insufficient to identify all individuals uniquely. To take the worst-case scenario—the most common male name with the most common birthplace—the 1880 U.S. census includes 17 white men aged 33 who were named John Smith and born in New York State. Even this example understates the problem, because it assumes an exact match of name and age. Errors in enumeration and transcription cause a significant proportion of matches to be imperfect: linking must be carried out on a probabilistic basis, allowing for imperfect correspondence of names and ages. Taking slight variations in age and name into account, there are about 50 potential matches for a 33 year-old John Smith born in New York State.

To reduce the potential for ambiguity, we follow the precedent of Ferrie (1996, 1999) and eliminate the most common names. In particular, we exclude from the linking universe "John Smith," and all other names that in combination with sex, birthplace and approximate age, identify more than one individual in the 1880 population. In the case of U.S. males, this restriction results in exclusion of 17.8 percent of the population for links between 1850 and 1880, 20.0 percent for 1860-1880, 20.6 percent for 1870-1880, and 22.4 percent for 1880-1900. The percentages rejected because of ambiguity are likely to be similar in Canada, but lower in Britain, Norway, and Iceland since those countries identify birthplace with far greater precision than does the U.S. or Canada.[4]

The exclusion of some names has the potential to introduce systematic selection bias. We will test the reduced set of names for representativeness with respect to ethnicity, occupational status, family relationship, place of residence, and other characteristics, and apply weights as necessary to minimize selection bias. Because we have representative individual-level data at the end of each linking interval, it is possible to estimate the characteristics of the population that survived between the censuses.[5] We can use these estimates to construct multidimensional weighting matrices based on the ratio of the linked population to the total population for linked persons with each combination of characteristics. When these matrices become too complex, we will turn to multi-stage weighting procedures of the sort used by the U.S. Census Bureau (2003).

Even though we focus on representativeness and accuracy instead of maximizing linkage rates, our strategy will result in far larger linked samples than have previously been available for nineteenth-century populations. For example, we estimate that we will link approximately 55,000 persons between 1870 and 1800 in the United States alone; altogether, we expect to produce some 650,000 linked cases. These large linked samples will sustain intensive analyses.

Name cleaning. Names are by far the most important piece of information available for record linkage, but they are also the most problematic. Errors in naming can arise from respondent error (as when, for example, a farm wife responding to an enumerator misstates the name of a farm hand), enumerator error, or transcription error. Moreover, names often change over time, sometimes did not have standard spellings, and in some cases people will be enumerated under a nickname or middle name in one census and under their formal first name in the other. To minimize error from these sources, we plan a program of name cleaning, accounting for common typographical transpositions, handwriting recognition errors, and common nicknames. Many of these techniques are language-specific and will have to be customized for each country. This work will draw on the rich body of research on name cleaning carried out during the past decade (Porter and Winkler 1997; Christen, Churches and Zhu 2002; Winkler 1990; Nygaard 1992, Maletic and Marcus 2000).

We will also employ phonetic name coding, a standard tool for record linkage since the 1930s. The most commonly used systems are Soundex, NYSIIS, and Phonex. All of these systems lose much of the phonetic

[4] A countervailing factor that could increase the percentage excluded, however, is that Iceland and Norway have less diversity of names than do the other countries.

[5] This technique will not work for forward linkages (e.g., 1870 to 1880) of immigrant populations, since we will not be able to distinguish immigrants who came during the linkage interval from those who were present previously. We will, however, be able to use demographic analysis to estimate with reasonable precision the total size of the persisting population of each age and sex for each immigrant group, and this will form the sole basis for the weights for linked immigrants. For backward linkages (e.g., 1900 to 1880) this is not a problem, since we know year of immigration and can therefore delineate the population that was present in both census years.

detail, however. Although we have not yet finalized our phonetic coding plans, we prefer the more subtle Double-Metaphone system, which returns two encoded strings corresponding to variant pronunciations (Philips 2000; Lait and Randell 1993). The multiple languages of the NAPP database complicate this problem, and we will have to take great care to ensure that any phonetic tools are appropriate for the particular language in which names are recorded.

Linking algorithm. Because there are multiple opportunities for errors to be introduced, it is essential that the linking algorithm accommodate approximate matches on a probabilistic basis. Planning and design of the linking algorithm is therefore a significant component of the project. The design must consider not only optimization of links, but also computational efficiency: some techniques are extraordinarily computationally intensive and would be unfeasible for a project of this scale (Christen et al. 2002).

The theoretical framework of record linkage derives from Fellegi and Sunter (1969), who demonstrated that it is possible to define an optimal linkage rule that minimizes the number of false links. In addition, Fellegi and Sunter derived a test statistic for evaluating error rates and specified the assumptions necessary for estimating the matching probabilities used to calculate the test statistic. Extensions and refinements of record-linkage theory were contributed by Jaro (1989), Winkler (1993), Belin and Rubin (1995), and Larson and Rubin (2001).

All these models assume that every pair of records drawn from two files are either matches referring to a single individual or non-matches describing two different persons; optimal matching requires that every individual be compared with every possible match. It is not computationally feasible to implement every potential match; for example, implementation of such a linking algorithm for the full U.S. 1880 database and the 1900 U.S. sample would involve over 15 trillion comparisons. To reduce the computational requirements, we will introduce "blocking factors"—such as state of birth and sex—and limit comparisons to persons who share the same blocking factors. If necessary, we will make an additional blocking pass based on metaphone. The computational problem will nevertheless be large, and we will carefully explore various methods that researchers have proposed to improve efficiency (e.g., Jin, Li, and Mehrotra 2003; Verykios, Elmagarmid and Houstis 2000; Hernádez and Stolfo 1998; Monge and Elkan 1997).

Because our linking strategy must rely heavily on names, identification of the optimal approximate string comparison algorithm is of paramount importance. Researchers have proposed many algorithms; based on our review of the literature, we presently favor the Jaro string comparator as modified by Winkler (Porter and Winkler 1997). This algorithm computes a similarity measure between 0.0 and 1.0 based on the number of common characters in two strings, the lengths of both strings, and the number of transpositions, accounting for the increased probability of typographical errors towards the end of words. Since developments in this field are proceeding rapidly, however, a superior algorithm may appear during the course of the project.

The other linking variables—such as birthplace, parental birthplaces, age, and sex—pose few string comparison problems because those variables are already classified and numerically coded according to the NAPP coding system. Thus, for example, we will not have to cope with the innumerable spelling variations of Reykjavík. We will, however, need to develop an algorithm for age misreporting that can account for digit preferences: inconsistencies in age between two census years should be partly discounted if age is rounded to a five or zero in one or both census years.

To estimate the matching parameters and error rate of the linking algorithm and to refine the linking strategy, we need a set of training data. Training data consist of cases where the true links are known. To estimate the parameters for imperfect matches on the key linking variables (e.g., name and age), we will turn to the married-couple linked samples; although these samples are not representative of the entire population, the matches will be highly accurate and will provide estimates of the effects of imprecision in predictors on the probability of incorrect matches. If necessary, we may also obtain training data by hand-coding a subset of cases using traditional genealogical methods.

Whenever possible, we plan to build on open-source software for both data cleaning and record linkage. In particular, we are excited by the potential of the "Freely extensible biomedical record linkage (Febrl)" software described in Christen and Churches (2002). We will also explore commercial software for cleaning and matching, such as Choicemaker, Trillium, and Matchmaker. Nevertheless, we recognize that we will have to develop the bulk of the cleaning and linking software in-house to meet our specialized needs.

Data access and online analysis. Data sharing is central to the project; effective dissemination is essential if the data are to be widely used. Both data and documentation will be distributed through an integrated web-based data access system. The IPUMS data access system—which is the basis of the current NAPP data access tool—pioneered web-based dissemination of large-scale datasets and has served as a model for many other social science data dissemination efforts. As part of the new IPUMS-International project, we are developing robust second-generation data dissemination software suitable for use across a wide range of datasets. The NAPP will leverage this investment by adapting the IPUMS-International software to the NAPP database.

Like the current software, this new data extraction system will allow users to merge datasets, select variables, and define population subsets. The new system will also offer advanced tools for navigating documentation, defining datasets, constructing customized variables, and adding contextual information. We will also add new data extraction tools to allow easy access to the linked files. These tools will offer users several options for rectangularizing the linked datasets at the individual or household level. The data extraction tool will incorporate documentation browsing and search functions so users have easy access to comprehensive documentation as they design their analyses.

We will also provide online data analysis for the cross-sectional datasets. The system will use an analysis engine developed by the Computer-assisted Survey Methods Program at the University of California, Berkeley. Survey Documentation and Analysis (SDA) uses an inverted-matrix data format and data packing techniques to deliver frequency distributions, cross-tabulations, means, and other simple statistical analyses for millions of observations in real time. The system also handles case selection and basic recoding. We expect that users will be able to tabulate the entire data series in less than a minute, making online analysis a feasible option for students and researchers without access to statistical software. As part of this project, we will develop a customized web interface for the SDA system tailored to the NAPP database. The availability of user-friendly online analysis will substantially broaden the audience for census microdata, and even sophisticated researchers will turn to the system for exploratory analyses.

The extraction engine is designed to take full advantage of the hierarchical structure of census data. We offer researchers the option of rectangular or hierarchical output files and allow users to select households or families based on individual-level characteristics. Future versions of the data access system will add additional features that exploit the hierarchical structure of the data by automating the creation of new variables describing the characteristics of subfamilies, families, and households.

Documentation and sustainability. The linked samples will require substantial new documentation. The documentation will include a full description of our procedures for creating the samples, estimates of error rates, and a usage manual with examples of appropriate analyses. We also plan to release all software used for the project, both for purposes of documentation and so that researchers can apply our cleaning and linking methods to other datasets.

All documentation will be compliant with the Data Documentation Initiative (DDI) XML metadata standard. The DDI was developed by an international committee that included the U.S. Census Bureau; the Bureau of Labor Statistics; the Inter-university Consortium for Political and Social Research (ICPSR); and the national data archives of Great Britain, Norway, and Canada (Block and Thomas 2003). The DDI is a non-proprietary, hardware independent, neutral standard that preserves the content and relational structure of the full documentation. The DDI was designed principally as an archival standard, but it also offers technical advantages. The machine-understandable structure of the DDI allows for automated processing by data-access software, and we are exploiting this capability in the new data access tools now under development. The documentation will be prepared exclusively in machine-readable form and will be disseminated through the Internet.

Long-run survival of the database beyond the project period is critical. DDI-encoded metadata will ensure that the microdata and documentation remain usable even if the technological environment shifts. The Minnesota Population Center, the Norwegian Historical Data Center, and the U.K. Data Archive all guarantee to maintain the system for a period of at least 25 years beyond the end of the project. We will also deposit the database and access software with ICPSR and the U.K. Data Archive to ensure permanent preservation.

In addition to the main user file consisting of numeric codes, we will also create an archive file. The archive file will consist of the unaltered transcription of all census items in alphabetic format. This will allow future investigators to construct alternate coding schemes and allocation procedures. The data dictionaries used to

translate information from alphabetic to numeric form will also be made available to the public and preserved by ICPSR and the U.K. Data Archive.

Collaborators

The investigators will work closely together, with continuous web-based interaction. We plan face-to-face meetings in the first, third, and fifth years of the project. If a consensus cannot be reached on particular design issues or coding decisions, the collaborators have all agreed to abide by the opinion of the majority. Steven Ruggles, Director of the Minnesota Population Center, is the Principal Investigator. He will be in charge of overall project coordination, will oversee software development, and will be in charge of the U.S. coding operation, which is funded by NICHD.

The international teams are located at the leading centers of population history infrastructure in each country, and represent an extraordinary pool of talent and expertise. Matthew Woollard, Head of the History Data Service of the U.K. Data Archive, will manage the British component of the project in consultation with Kevin Schürer, Director of the U.K. Data Archive. Lisa Y. Dillon, Assistant Professor of Historical Demography at the Université de Montréal and President of the International Microdata Access Group will direct the Canadian effort, assisted by Chad Gaffield, Director of the Institute for Canadian Studies at the University of Ottawa and Principal Investigator of the Canadian Century Infrastructure project. The Norwegian project will be jointly headed by Gunnar Thorvaldsen, Director of the Norwegian Historical Data Centre and Jan Oldervoll, Co-Director of the Digital Archive of the Norwegian National Censuses. Finally, Ólöf Garðarsdóttir, Head of Demographic Statistics at Statistics Iceland, will be in charge of the Icelandic component of the project.

Professor Joseph Ferrie of Northwestern University will assist with national and international linkage of historical census records. Ferrie has already experimented with automatic linkage of NAPP data to sample data within and between Britain and the United States, and his experience will be an invaluable asset to the project (Ferrie 2003). The project coordinator is Evan Roberts, a doctoral candidate in economic history at the University of Minnesota. Roberts has proven to be a talented diplomat and negotiator, capable of finding a consensus among the eight researchers about every aspect of the database. Fortunately, he has agreed to continue serving the project on a part-time basis after he completes his dissertation.

The integration of new datasets will proceed in the same fashion as in the first phase of NAPP. Each research center will be responsible for cleaning, reformatting, and coding their own datasets, and we will trade batches of strings to allow blind verification of coding decisions. The linking component of the project will also be a fully collaborative effort. The Minnesota Population Center will have lead responsibility for software development, but will work in close consultation with each of the other research centers. Each center will develop training datasets of hand-linked cases needed to tune the linking algorithm to local conditions. Name-cleaning routines—which standardize common abbreviations, nicknames, and misspellings—are dependent on language and custom, and will be developed separately in each country. The U.K. Data Archive will compare the characteristics of the automatically linked samples with one compiled entirely by genealogical methods, and all centers will evaluate the reliability of a sample of links by checking them manually. The entire database will be distributed by the Norwegian Historical Data Centre, the U.K. Data Archive, the Université de Montréal program in Historical Demography, and the Minnesota Population Center.

Schedule of work and deliverables

If this project is funded now, work can begin immediately with the same staff as the existing project. In the first year, programmers at Minnesota will standardize the format of the datasets to be added to NAPP, identify character strings that need to be coded, and add the uncoded strings to the variable dictionary management system. The dictionaries will then be redistributed to each participating country, and we will begin the production phase of the data dictionary work. We will release new datasets as the cleaning and dictionary work is completed; most of the new datasets will be available to researchers by May 2008, but we anticipate that a few datasets from Norway, Iceland, and Canada will not be complete until the end of the project, since data entry is still taking place. The linking project for each country will begin as soon as preliminary datasets are ready. The Norwegian and British teams will prepare training datasets by March 2007, and we will implement the linking algorithm in those countries during the following year. The Canadian and Icelandic linked samples will follow in 2009. We will release all final linked samples, including the internationally-linked samples, in 2010. The first version of the new data access software will be released by September 2005; thereafter, we plan new versions on an annual basis. Online data analysis will be incorporated into the system by September 2007.

References

Anderson, Michael. 1972. *Family Structure in Nineteenth Century Lancashire.* Cambridge, England: Cambridge University Press.

Anderson, Michael., Stott, C., and Collins, B. 1979. National Sample from the 1851 Census of Great Britain [computer file]. Colchester, Essex: UK Data Archive [distributor].

Belin, T.R. and D.B. Rubin. 1995. A Method for Calibrating False-match Rates in Record Linkage. *Journal of the American Statistical Association* 90: 694-707.

Block, William and Wendy Thomas. 2003. Implementing the Data Documentation Initiative at the Minnesota Population Center. *Historical Methods* 36 (2): 97-101.

Cho, Lee-Jay; Retherford, Robert D.; Choe, Minja Kim. 1986. *The Own-Children Method of Fertility Estimation.* Honolulu: University of Hawaii Press.

Christen, Peter and Tim Churches. 2002. Febrl - Freely extensible biomedical record linkage. *ANU Computer Science Technical Reports TR-CS-02-05*, Australian National University, Canberra, October 2002. <http://datamining.anu.edu.au/software/>

Christen, Peter, Tim Churches and Justin Xi Zhu. 2002. Probabilistic Name and Address Cleaning and Standardisation. *Proceedings of the Australasian Data Mining Workshop*, December, Canberra, Australia.

Christen, Peter, Justin Xi Zhu, Markus Hegland, Stephen Roberts, Ole M. Nielsen, Tim Churches and Kim Lim. 2002. High-Performance Computing Techniques for Record Linkage. Proceedings of the Australian Health Outcomes Conference (AHOC-2002), Canberra, Australia.

Coale, Ansley J. and Susan C. Watkins, eds. 1986. *The Decline of Fertility in Europe*. Princeton: Princeton University Press.

Committee on Applied and Theoretical Statistics, National Research Council. 1999. *Record Linkage Techniques - 1997: Proceedings of an International Workshop and Exposition* Washington, D.C.: National Academy Press.)

Edvinsson, Sören and Johnny Karlsson. 1998. Recoding occupations in the Demographic Data Base into HISCO. In *HISMA Occasional Papers & Documents Series*, Berlin. No. 3/1998.

Fellegi, I. P., and A.B. Sunter. 1969. A Theory for Record Linkage, *Journal of the American Statistical Association*, 40: 1183-1210.

Ferrie, Joseph. 1996. A New Sample of Males Linked from the Public-Use-Microdata-Sample of the 1850 US Federal Census of Population to the 1860 US Federal Census Manuscript Schedules. *Historical Methods* 29: 141-156.

Ferrie, Joseph. 1999. How Ya Gonna Keep 'Em Down on the Farm [When They've Seen Schenechtady]?: Rural to Urban Migration in Nineteenth-Century America 1850-1870. Working paper, Northwestern Univ. <http://www.faculty.econ.northwestern.edu/faculty/ferrie/papers/urban.pdf>

Ferrie, Joseph P. 2003. Longitudinal Data for the Analysis of Mobility in the U.S., 1850-1930. Paper delivered at Longitudinal and Cross-Sectional Historical Data: Intersections and Opportunities, IMAG, Montreal, Nov. 10-11, 2003 <http://www.nappdata.org/imagpapers/ferrie.pdf>

Ferrie Joeseph P. and Jason Long. 2004. A Tale of Two Labor Markets: Career Mobility in Britain (1851-81) and the U.S. (1850-80). Presented at Allied Social Science Association Meetings (January 2004) and European Social Science History Conference (Berlin, March 2004). <http://www.colby.edu/economics/faculty/jmlong/research/usukmobility.pdf>

Foster, John O. 1974. *Class Struggle in the Industrial Revolution: Early Industrial Capitalism in Three English Towns*. London.

Gjerde, Jon. 1992. *From Peasants to Farmers: the Migration from Balestrand, Norway to the Upper Middle West.* Cambridge: Cambridge University Press.

Hacker, J. David. 1999. The Human Cost of War: White Population in the United States, 1850-1880. Ph.D. dissertation, University of Minnesota.

Hacker, J. David. 2000a. Rethinking the 'Early' Decline of American Fertility. Paper presented at the Minnesota Population Seminar, October 2000.

Hacker, J. David. 2000b. New estimates of census underenumeration in the United States, 1850-1880. Paper presented at the annual meeting of the Population Association of America, Los Angeles.

Hacker, J. David. 2003. Rethinking the 'Early' Decline of Marital Fertility in the United States. *Demography* 40 (4): 605-620.

Hall, Patricia Kelly, Robert McCaa, and Gunnar Thorvaldsen (eds). 2000. *Handbook of International Historical Microdata for Population Research*. Minneapolis: Minnesota Population Center. <http://www.ipums.org/international/microdata_handbook.html>

Hareven, Tamara K. 1978. The dynamics of kin in an industrial community. In *Turning Points: Historical and Sociological Essays on the Family* John Demos and S.S. Boocock, eds. Chicago: University of Chicago Press.

Hareven, Tamara K. 1982. *Family Time and Industrial Time: The Relationship between the Family and Work in a New England Industrial Community*. New York: Cambridge University Press.

Hernádez, M.A. and S.J. Stolfo. 1998. Real-World Data is Dirty: Data Cleansing and the Merge/purge Problem. *Data Mining and Knowledge Discovery* 2 (1):9-37. <http://www.cs.columbia.edu/~sal/hpapers/mp.ps>

Jaro, Matthew A. 1989. Advances in Record Linking Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association* 84: 414-20.

Jin, Liang, Chen Li, and Sharad Mehrotra. 2003. Efficient Record Linkage in Large Data Sets. Presented at the 8th International Conference on Database Systems for Advanced Applications (DASFAA 2003) March, Kyoto, Japan. <http://www.ics.uci.edu/~chenli/pub/dasfaa03.pdf>

Katz, Michael B. 1975. *The people of Hamilton, Canada West: family and class in a mid-nineteenth-century city*. Cambridge: Harvard University Press.

King, Miriam L. and Diana L. Magnuson. 1995. Perspectives on historical U.S. census undercounts. *Social Science History* 19(4): 455-66.

Knights, Peter R. 1991. *Yankee Destinies: the Lives of Ordinary Nineteenth-Century Bostonians*. Chapel Hill: University of North Carolina Press.

Lait, A.J. and B. Randell. 1993. An Assessment of Name Matching Algorithms. Department Technical Report Series No. 550, Department of Computing Science, University of Newcastle upon Tyne, UK. <http://homepages.cs.ncl.ac.uk/brian.randell/home.informal/Genealogy/NameMatching.pdf>

Larson, M. and D.B. Rubin. 2001. Iterative Automated Record Linkage Using Mixture Models. *Journal of the American Statistical Association* 96: 32-41

Maletic, Jason I. and Andrian Marcus. 2000. Data cleansing: Beyond Integrity Analysis. In *Proceedings of the Conference on Information Quality*, (Boston, MA: Massachusetts Institute of Technology) October 20-22, 2000, pp. 200-209.

Modell, John. 1978. Patterns of consumption, acculturation, and family income strategies in late nineteenth-century America. In T.K. Hareven and M. Vinovskis, eds. *Family and Population in Nineteenth Century America.* Princeton: Princeton University Press.

Monge, Alvaro E. and Charles Elkan. 1997. An Efficient Domain-independent Algorithm for Detecting Approximately Duplicate Database Records. Presented at SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'97), May Tucson, Arizona.

Nugent, Walter. 1992. *Crossings: the Great Transatlantic Migrations, 1870-1914*. Bloomington, Indiana: Indiana University Press.

Nygaard, Lars. 1992. Name Standardization in Record Linkage: an Improved Algorithmic Strategy. *History and Computing* 4(2): 63-74.

Oldervoll, Jan. 1982. Maskinell kopling av eit historisk materiale (Automated linkage of a historic source material). Paper presented at the seminar Historical Databases in the Nordic Countries, Sandbjerg, Denmark, 15-18 September.

Perlmann, Joel. 2003. IPUMS. (Web Site Review) *Journal of American History* June 2003 (Vol. 90, No. 1) pp. 339-340. <http://ipums.org/jah.html>

Philips, Lawrence. 2000. The Double-Metaphone Search Algorithm, *C/C++ User's Journal* 18(6).

Porter, Edward H. and William E. Winkler. 1997. Approximate String Comparison and its Effect on an Advanced Record Linkage System. Census Bureau Research Report RR97/02 (Washington D.C.: U.S. Bureau of the Census). <http://www.fcsm.gov/working-papers/porter-winkler.pdf>

Roberts, Evan, Steven Ruggles, Lisa Dillon, Olof Gardarsdottir, Jan Oldervoll, Gunnar Thorvaldsen, and Matthew Woollard. 2003a. The North Atlantic Population Project: An Overview. *Historical Methods* 36 (2): 80-88.

Roberts, Evan, Matthew Woollard, Chad Ronnander, Lisa Dillon, and Gunnar Thorvaldsen. 2003b. Occupational Classification in the North Atlantic Population Project. *Historical Methods* 36 (1): 89-96.

Rosenthal, Paul-Andre. 1997. Thirteen Years of Debate: From Population History to French Historical Demography (1945-1958). *Population: An English Selection* 9: 215-241.

Ruggles, Steven. 1994. The transformation of American family structure. *American Historical Review* 99: 103-28.

Ruggles, Steven. 2003a. Multigenerational Families in Nineteenth-Century America. *Continuity And Change* 18 (1): 139-165.

Ruggles, Steven. 2003b. Linking Historical Censuses: A New Approach. Paper delivered at Longitudinal and Cross-Sectional Historical Data: Intersections and Opportunities, International Microdata Access Group, Montreal, Nov. 10-11, 2003 <http://www.nappdata.org/imagpapers/ruggles.pdf>

Runblom, Harald and Hans Norman, eds. 1976. *From Sweden to America: a history of the migration*. Minneapolis, Minnesota: University of Minnesota Press.

Sacerdote, Bruce. Forthcoming. 2005. Slavery and the Intergenerational Transmission of Human Capital. *The Review of Economics and Statistics*. 87(2).

Thernstrom, Stephan. 1964. *Poverty and progress; social mobility in a nineteenth century city* Cambridge, Massachusetts: Harvard University Press.

Thorvaldsen, Gunnar. 1995. *Migration in the province of Troms 1865-1900. A Study based on the Censuses.* (Dr. philos avhandling: *Migrasjon i Troms i annen halvdel av 1800-tallet. En kvantitativ analyse av folketellingene 1865, 1875 og 1900)*.

Thorvaldsen, Gunnar. 1997. Marriage and names among immigrants to Minnesota, *Electronic Journal of the the American Association for History and Computing.* <http://mcel.pacificu.edu/history/JAHC/Thorvaldsen/ThorIndex.html>

van Leeuwen, H.D., Ineke Maas and Andrew Miles. 2002. *HISCO: Hostorical International Standard Classification of Occupations.* Leuven: Leuven University Press

Verykios, Vassilios S., Elmagarmid, A.K., and Houstis, E.N. 2000. Automating the Approximate Record Matching Process. *Journal of Information Sciences* 126 (1-4): 83-98.

Winkler, William E. 1990. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *American Statistical Association 1990 Proceedings of the Section of Survey Research Methods*, 354-359. <http://www.amstat.org/sections/srms/Proceedings/papers/1990_056.pdf>

Winkler, William E. 1993. Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage. *American Statistical Association 1993 Proceedings of the Section of Survey Research Methods*, 274-279. <http://www.census.gov/srd/papers/pdf/rr93-12.pdf>

Biosketch
Steven Ruggles

Minnesota Population Center
University of Minnesota, Twin Cities
537 Heller Hall, 271 19th Avenue South
Minneapolis, MN 55455-0406

Phone: (612) 624-5818
Fax: (612) 626-8375
E-mail: ruggles@pop.umn.edu
Web: http://www.pop.umn.edu/~ruggles

Professional Preparation

University of Wisconsin	History	B.A., 1978
University of Pennsylvania	History	M.A., 1982
University of Pennsylvania	History	Ph.D., 1984
University of Wisconsin	Sociology/Demography	NICHD Post-Doctoral Trainee, 1984-85

Appointments

2000-	Distinguished McKnight University Professor, University of Minnesota Director, Minnesota Population Center, University of Minnesota
1995-	Professor of History; graduate faculty in Sociology, Public Policy, Population Studies, and American Studies, University of Minnesota
1988-1994	Associate Professor of History, University of Minnesota
1985-1987	Assistant Professor of History, University of Minnesota

Publications Related to Proposal

S. Ruggles and J.D. Hacker (eds). 2003. *Building Infrastructure for the Social Sciences*. Two Special Issues, *Historical Methods.* 36. (1-2): 5-101.

S. Ruggles and S. Brower. 2003. "The Measurement of Family and Household Composition in the United States, 1850-1999." *Population and Development Review.* 29(1): 73-101.

R. McCaa and S. Ruggles. 2002. "The Census in Global Perspective and the Coming Microdata Revolution." *Scandinavian Population Studies.* 13: 7-30.

S. Ruggles, and P. Kelly Hall, eds. 1999. *IPUMS: The Integrated Public Use Microdata Series*. Special issue, *Historical Methods.* 32(3): 102-158.

S. Ruggles, M. Sobek and T. Gardner. 1996. "Distributing Large Historical Census Samples on the Internet." *History and Computing.* 8(3): 145-59.

Other Significant Publications

S. Ruggles. 1997. "The Rise of Divorce and Separation in the United States, 1880-1990." *Demography.* 34(4): 455-66.

S. Ruggles. 1994. "The Transformation of American Family Structure." *American Historical Review.* 99(1):103-28.

S. Ruggles. 1994. "The Origins of African-American Family Structure." *American Sociological Review.* 59(1): 136-51.

S. Ruggles. 1992. "Migration, Marriage, and Mortality: Correcting Sources of Bias in English Family Reconstitutions." *Population Studies.* 46(3): 507-522.

S. Ruggles. 1987. *Prolonged Connections: The Rise of the Extended Family in Nineteenth Century England and America.* Madison: University of Wisconsin Press.

Synergistic Activities

1. Principal investigator or co-principal investigator on 21 large social science infrastructure projects devoted to data improvement, harmonization, and interoperability. These projects, with total costs of $37 million, include the "Integrated Public Use Microdata Series." (IPUMS-USA), "International Integrated Microdata Access system." (IPUMS-International), "North Atlantic Population Project." and "National Historical Geographic Information System." These efforts resulted in some of the most widely used databases for research on human and social dynamics, and in 2003 the Population Association of America recognized Ruggles' contributions to social science infrastructure with the Robert J. Lapham award.
2. Co-founder of the Demographic Data Cooperative, an organization dedicated to promoting interoperability among population research center data collections and developing electronic data dissemination tools and instructional materials.
3. Council member, Inter-university Consortium for Political and Social Research, and chair of the Archival Development Committee and the Subcommittee on Census 2000.
4. Chair, Task Force on Census 2000.

Collaborators within the last 48 months

John S. Adams	University of Minnesota	Miriam King	University of Minnesota
Andrew Beveridge	Queens College, CUNY	Felicia LeClere	Notre Dame University
Lynn A. Blewett	University of Minnesota	Deborah Levison	University of Minnesota
William C. Block	University of Minnesota	Carolyn Liebler	University of Minnesota
Michael Davern	University of Minnesota	Robert McCaa	University of Minnesota
Lisa Y. Dillon	University of Montreal	Robert B. McMaster	University of Minnesota
Catherine A. Fitch	University of Minnesota	Russell Menard	University of Minnesota
William Finzer	Key Curriculum Press	Cuong Nguyen	University of Minnesota
Chad Gaffield	University of Ottawa	Jan Oldervoll	University of Bergen
Todd Gardner	U.S. Census Bureau	Alberto Palloni	University of Wisconsin
Ólöf Garðarsdóttir	Statistics Iceland	Evan Roberts	University of Minnesota
Ron Goeken	University of Minnesota	Walter Sargent	University of Minnesota
Myron P. Gutmann	University of Michigan	Kevin Schurer	UK Data Archive
J. David Hacker	SUNY-Binghamton	Matthew Sobek	University of Minnesota
Michael Haines	Colgate University	James A. Sweet	University of Wisconsin
Patricia Kelly Hall	University of Minnesota	Wendy Thomas	University of Minnesota
Jeremy Holzman	MN Medical Research Foundation	Gunnar Thorvaldsen	University of Tromsø
		Stewart Tolnay	Univ. of Washington
Dirk Jaspers	CELADE	Matthew Woollard	University of Essex

Graduate and Post Doctoral Advisors

Michael B. Katz	University of Pennsylvania	James A. Sweet	University of Wisconsin

Thesis Advisor and Postgraduate-Scholar Sponsor (18 total)

Deborah Alexander	University of Minnesota	Patricia Kelly Hall	University of Minnesota
Trent Alexander	University of Minnesota	Daniel C. Kallgren	University of Wisconsin
William C. Block	University of Minnesota	Miriam L. King	University of Minnesota
Jason Digman	University of Minnesota	Carolyn Liebler	University of Minnesota
Lisa Y. Dillon	University of Montreal	Diana L. Magnuson	Bethel College
Catherine A. Fitch	University of Minnesota	Agnes Odinga	Hamline University
Todd Gardner	U. S. Census Bureau	Evan Roberts	University of Minnesota
Ron Goeken	University of Minnesota	Chad Ronnander	University of Minnesota
J. David Hacker	SUNY-Binghamton	Matthew Sobek	University of Minnesota

Biosketch
Lisa Y. Dillon

Département de Démographie, Université de Montréal
C.P. 6128, succursale Centre-ville
Montréal, Québec, H3C 3J7
Canada

Phone: (514) 343-5956
Fax: (514) 343-2309
E-mail: ly.dillon@umontreal.ca

Professional Preparation

University of Waterloo	History/English	B.A., 1985
University of Ottawa	History	M.A., 1990
University of Minnesota	History	Ph.D., 1997
University of Victoria	History	Canadian Families Project, Post-Doctoral Fellow, 1998
University of Ottawa	Canadian Studies	SSHRC Post-Doctoral Fellow, 1998-2000

Appointments

2001-	Assistant Professor, Département de Démographie, Université de Montréal, Montréal, Québec.
2000-2001	Researcher and Research Co-ordinator, Institute of Canadian Studies, University of Ottawa

Publications Related to Proposal

K. Mandemakers and L. Dillon. 2004. "Best practices with large databases on historical populations." *Historical Methods*. 37(1): 34-39.

E. Roberts, M.Woollard, C. Ronnander, L. Dillon and G. Thorvaldsen. 2002. "Occupational Classification in the the North Atlantic Population Project." *Historical Methods.* 36(2): 89-96.

E. Roberts, S. Ruggles, L. Dillon, Ó. Garðarsdóttir, J. Oldervoll, G. Thorvaldsen and M. Woollard. 2002. "The North Atlantic Population Project: An Overview." *Historical Methods*. 36(2): 80-88.

L. Dillon. 2000. "International Partners, Local Volunteers and Lots of Data: The 1881 Canadian Census Project." *History and Computing.* 12(2): 163-176.

L. Dillon. 2000. "Integrating Canadian and U.S. historical census microdata: 1871 and 1901 Canada, and 1870 and 1900 United States." *Historical Methods*. 33(4): 185-194.

Other Significant Publications

L. Dillon. (forthcoming). "Boundaries of Age: Exploring the patterns of young-old age among men, Canada and the United States, 1870-1901." *Canada's Families at the Turn of the Twentieth Century*, (Toronto: University of Toronto Press).

L. Dillon and B. Desjardins. (forthcoming). "La définition de la vie familiale en Nouvelle-France : les seuils de la vieillesse dans une optique familiale, » dans Patrice Vimard et Kokou Vignikin." *Familles au Nord, Familles au Sud*. (Academia-Bruylant/L'Harmattan).

L. Dillon. 2001. "Elderly Women in Late Victorian Canada." In Bob Hesketh and Chris Hackett (eds). *Canada: Confederation to Present*. (CD-ROM). (Edmonton: University of Alberta).

L. Dillon. 1999. "Women and the Dynamics of Household Headship, Marriage and Aging in Victorian Canada and the United States." In E. Sager and P. Baskerville, eds. *Family History, An International Quarterly: Special Issue on Canadian Family History.* 4(4): 447-483.

L. Dillon. 1998. "Parent-Child Co-Residence Among the Elderly in Victorian Canada and the United States: A Comparative Study." In E.-A. Montigny and L. Chambers (eds). *Family Matters: Papers in Post-Confederation Canadian Family History*. (Toronto: Canadian Scholars Press).

Synergistic Activities

1. Creator of research program at Université de Montréal that harmonizes and disseminates longitudinal and cross-sectional Québec and Canadian population data from the 17th to the 19th centuries. Funding for the program comes from three major research grants from Social Sciences and Humanities Research Council of Canada, la Fondation québécoise de recherche en sciences et cultures, and the Canadian Foundation for Innovation. Dillon established the Historical Demography Research Infrastructure with a total budget of $601,166. This research laboratory and program of database construction and research employs ten people.

2. Chair (1998-present), International Microdata Access Group (IMAG) to foster international collaboration of interdisciplinary researchers working with individual-level population data to facilitate transnational comparative research. Organized second IMAG international workshop "Longitudinal and Cross-Sectional Historical Data: Intersections and Opportunities" held at the Université de Montréal November 10-11, 2003 and featuring scholars from the U.S., the Netherlands, Sweden, Belgium and Switzerland. Conference supported by a Social Sciences and Humanities Research Council of Canada conference grant ($10,000).

3. Presented research findings at international conferences, including the Journées Scientifiques de l'Association des démographes, the European Social Science History Association and the International Association of History and Computing. Invited talks at: University of Iceland, University of Ottawa and York University.

4. Principal Investigator (2000-2001), the 1881 Canadian Census Project, University of Ottawa Research Partnerships Programme and Church of Jesus Christ of Latter-day Saints ($34,658).

5. Principal Investigator (1998-2000), Social Sciences and Humanities Research Council Strategic Research Development Initiatives Grant which supported the first IMAG workshop at the University of Ottawa, 1999 ($50,000).

Collaborators within the last 48 months

Bertrand Desjardins	Université de Montréal	Ólöf Gardarsdottir	Statistics Iceland
Kees Mandemakers	International Institute of Social History	Jan Oldervoll	University of Bergen
		Matthew Woollard	University of Essex
Steven Ruggles	University of Minnesota	Kevin Schürer	University of Essex
Evan Roberts	University of Minnesota	Gunnar Thorvaldsen	University of Tromsø
Chad Ronnander	University of Minnesota		

Graduate and Post Doctoral Advisors

Steven Ruggles	University of Minnesota	Chad Gaffield	University of Ottawa

Thesis Advisor and Postgraduate-Scholar Sponsor

Edouard Nakouzi, Université de Montréal
Josiane Doucet-Alarie, Université de Montréal
Stéphanie Villeneuve, l'Université Marc Bloc de Strasbourg

Biosketch
Joseph P. Ferrie

Department of Economics
Northwestern University
302 Arthur Andersen Hall, 2001 Sheridan Road
Evanston, IL 60208-2600
Web: http://www.faculty.econ.northwestern.edu/faculty/ferrie/

Phone: (847) 491-8210
Fax: (847) 491-7001
E-mail: ferrie@northwestern.edu

Professional Preparation

Williams College	History and Economics	B.A., 1983
University of Chicago	Economics	M.A., 1987
University of Chicago	Economics	Ph.D., 1992

Appointments

2004	Visiting Scholar, Institut National de la Recherche Agronomique, Paris, France
2000-	Faculty Associate, Institute for Policy Research, Northwestern University
1997-	Associate Professor of Economics, Northwestern University
1997-	Research Associate, National Bureau of Economic Research
1991-2000	Faculty Fellow, Institute for Policy Research, Northwestern University
1991-1997	Assistant Professor of Economics, Northwestern University
1991-1997	Faculty Research Fellow, National Bureau of Economic Research

Publications Related to Proposal

J.P. Ferrie. 2003. "The Rich and the Dead: Socioeconomic Status and Mortality in the U.S., 1850-1860." In Dora Costa (ed). *Health and Labor Force Participation Over the Life Cycle*. (Chicago: University of Chicago Press).

J.P. Ferrie. 1999. *Yankeys Now: European Immigrants in the Antebellum U.S., 1840-60*. Oxford: Oxford University Press.

J.P. Ferrie. 1997. "The Entry into the U.S. Labor Market of Antebellum European Immigrants, 1840-1860." *Explorations in Economic History.* 34(3): 295-330.

J.P. Ferrie. 1997. "Immigrants and Natives: Comparative Economic Performance in the United States, 1850-1860 and 1965-1980." *Research in Labor Economics.* 16: 319-341.

J.P. Ferrie. 1996. "A New Sample of Americans Linked From the 1850 Public Use Micro Sample of the Federal Census of Population to the 1860 Federal Census Manuscript Schedules." *Historical Methods*. 29(4): 141-156.

Other Significant Publications

J.M. Currie and J.P. Ferrie. 2000. "The Law and Labor Strife in the U.S., 1881-1894." *Journal of Economic History*. 60(1): 1-25.

L.J. Alston and J.P. Ferrie. 1999. *Southern Paternalism and the Rise of the Welfare State: Economics, Politics, and Institutions in theU.S. South, 1865-1965.* New York: Cambridge University Press.

L.J. Alston and J.P. Ferrie. 1993. "Paternalism in Agricultural Labor Contracts in the U.S. South: Implications for the Growth of the Welfare State." *American Economic Review*. 83(4): 852-876.

L.J. Alston and J.P. Ferrie. 1989. "Social Control and Labor Relations in the American South Before the Mechanization of the Cotton Harvest in the 1950s." *Journal of Institutional and Theoretical Economics.* 145(1): 133-157.

L.J. Alston and J.P. Ferrie. 1985. "Resisting the Welfare State: Southern Opposition to the Farm Security Administration." In Robert Higgs (ed). *Research in Economic History* Supplement 4: Emergence of the Modern Political Economy: 83-120.

Synergistic Activities

1. Principal investigator or co-principal investigator on two current NSF funded projects (SES-0111838 and SES-0112093). These projects use U.S. census data to examine socio-economic status and mortality in the United States between 1850 and 1880 and the extent of social mobility among American farmers in the early twentieth century.
2. Principal investigator on three previous NSF funded grants that created panel data sets for individuals from historical U.S. censuses
3. Consultant for Early Indicators Project at Center for Population Economics, University of Chicago.
4. Consultant for NIH funded Alameda County Health Study at University of Michigan, studying feasibility of linking sample subjects to pre-1940 U.S. census manuscripts.

Collaborators within the last 48 months

Lee J. Alston, University of Colorado
Jason Long, Colby College
Steven Ruggles, University of Minnesota
Werner Troesken, University of Pittsburgh

Graduate and Post Doctoral Advisors

David W. Galenson, University of Chicago
Robert W. Fogel, University of Chicago

Thesis advisor and Postgraduate-Scholar Sponsor

Stewart, James Ireland (Reed College). "Essays on the economic history of the American frontier." Northwestern University, 2003. Thesis advisor.

Long, Jason (Colby College). "Labor mobility in Victorian Britain." Northwestern University, 2002. Thesis advisor.

Biosketch
Chad Gaffield

Institute of Canadian Studies
University of Ottawa
52 University Street
Ottawa, Ontario
Canada K1N 6N5

Phone: (613) 562-5111
Fax: (613) 562-5216
E-mail: gaffield@uottawa.ca

Web: http://www.canada.uottawa.ca/ccri

Professional Preparation

McGill University	History	B.A., 1973
McGill University	History	M.A., 1974
University of Toronto	History of Education	Ph.D., 1978

Appointments

2003-	University Research Chair, University of Ottawa
1997-2003	Founding Director of the Institute of Canadian Studies, University of Ottawa
1990-1996	Vice-Dean, Graduate Studies and Research, University of Ottawa
1988-	Full Professor, Department of History, University of Ottawa
1979-1985	Assistant/Associate Professor, University of Victoria, British Columbia
1978-1979	Visiting Assistant Professor, McGill University

Publications Related to Proposal

C. Gaffield (ed). 2003. *The Canadian Distinctiveness into the Twenty-first Century/La distinction canadienne au XXIième siècle* Ottawa: University of Ottawa Press.

C. Gaffield. 2002. "Inferring and Revising Theories with Confidence: Data Mining the 1901 Canadian Census." in *Mining Official Data, ECML/PKDD.* Helsinki.

C. Gaffield. 2002. "Historical Thinking, C.P. Snow's Two Cultures, and a Hope for the 21st Century." *Journal of the Canadian Historical Association*,*Quebec 2001.* (Ottawa: CHA): 3-25.

C. Gaffield. 2000. "Linearity, Non-Linearity, and the Competing Constructions of Social Hierarchy in Early Twentieth-Century Canada: The Question of Language in 1901." *Historical Methods* 33(4): 255-260.

C. Gaffield. 2000. "Time, Mathematics and Mass Schooling in 19th Century Ontario and Quebec." in Hubert Watelet, ed., *Quatre Essais sur Temps et Culture*. (Quebec: CIEQ): 33-39.

Other Significant Publications

C. Gaffield (ed). 2000. *The Doukhobor Centenary in Canada*. Ottawa: Slavic Research Group at the University of Ottawa & Institute of Canadian Studies, University of Ottawa.

C. Gaffield (ed). 1995. *Consuming Canada: Selected Reading in Environmental History* Toronto: Copp Clark Longman.

C. Gaffield (ed). 1994. *Constructing Modern Canada: Readings in Postconfederation History.* Toronto: Copp Clark Longman.

C. Gaffield. 1987. *Language, Schooling, and Cultural Conflict: The Origins of the French-language Controversy in Ontario*. Montreal, Kingston: McGill-Queen's University Press.

C. Gaffield. 1982. "Theory and Method in Canadian Historical Demography." *Archivaria*. 14: 123-136.

Synergistic Activities

1. Principal investigator or co-principal investigator on 6 large historical research projects focused on nineteenth and twentieth century social, economic, demographic and cultural change. These projects include The Canadian Social History Project (1975-1978); the Vancouver Island Project (1982-1985); the Historical Atlas of Canada Project, Vol.3 (1981-1988); The Lower Manhattan Project (1983-1990); The Outaouais Regional History Project (1988-1994); the Canadian Families Project (1995-2000); and the Canadian Century Research Infrastructure Project (2003-2008). These projects, with total costs of $19 million, resulted in publications and research/teaching infrastructure that has been recognized by awards given by the Canadian Historical Association, the Ontario Historical Association, and the Royal Society of Canada as well as by the Government of Canada in the form of the Queen's Jubilee Medal.
2. President of learned societies such as the Humanities and Social Sciences Federation of Canada (representing 25,000 researchers), the Canadian Historical Association, and the Canadian Historical of Education Association.
3. Co-founder of the Data Liberation Initiative, an effort that transformed the dissemination of research data for educational and academic purposes, and thereby launched a new era in Canadian data dissemination.
4. Council member of many organizations including the Council of Ontario Universities, the Association for Canadian Studies (both in Canada and in the United States), the National Capital Commission, and the External Advisory Committee (Statistics Canada).

Collaborators within the last 48 months

Gordon Darroch	York University	Patricia Kelly Hall	University of Minnesota
Peter Baskerville	University of Victoria	Robert McCaa	University of Minnesota
Claude Bellavance	University of Quebec	Jan Oldervoll	University of Bergen
Marc St-Hilaire	Laval University	Evan Roberts	University of Minnesota
Sean Cadigan	Memorial University	Walter Sargent	University of Minnesota
Lorne Tepperman	University of Toronto	Kevin Schurer	UK Data Archive
Charles Jones	University of Toronto	Matthew Sobek	University of Minnesota
Carl Amrhein	University of Alberta	Wendy Thomas	University of Minnesota
Lisa Y. Dillon	University of Montreal	Gunnar Thorvaldsen	University of Tromsø
Ólöf Garðarsdóttir	Statistics Iceland	Matthew Woollard	University of Essex
Ron Goeken	University of Minnesota		

Graduate and Post Doctoral Advisors

John Hellman	McGill University	Ian Winchester	University of Toronto

Thesis Advisor and Postgraduate-Scholar Sponsor

(total advisees: 28 master's theses, 11 doctoral theses, 3 post-docs)
Selected former doctoral and post-doctoral students and current positions:

John Lutz	Associate Professor, University of Ottawa
Lorne Hammond	Archivist, Public Archives of British Columbia
Barbara Lorenzkowski	Assistant Professor, Nippissing University
Christopher Clarkson	Assistant Professor, Cariboo College, British Columbia
John Bonnett	Research Scientist, National Research Council of Canada
Jo-Anne McCutcheon	President, CDCI
Lisa Dillon	Assistant Professor, University of Montreal
Eileen O'Connor	Assistant Professor, University of Ottawa

Biosketch
Ólöf Garðarsdóttir

Statistics Iceland, Population department
Borgartún 21a, 150 Reykjavík
Iceland
Email: olof.gardarsdottir@statice.is

Phone: (+354) 528-1031
(+354) 553-3846
(+354) 868-9772
Fax: (+354) 528-9911

Professional Preparation

University of Umeå, Sweden	History	Ph.D.	2002
University of Iceland, Reykjavik, Iceland	History	M.A.	1995
University of Iceland, Reykjavik, Iceland	History	B.A.	1993
University of Education, Reykavik, Iceland	Teacher's program	B.Ed.	1987

Appointments

2002-	Head of population statistics, Statistics Iceland, Reykjavík.
2003-2004	University of Education, Reykjavík, Iceland. Course on social history and demography.
2002-2003	University of Iceland, Reykjavík. Course on social history and history of economics.
2002-2003	Instructor in various courses at the University of Iceland
2001-2002	University of Umeå, Sweden. Course on family history and demography.
1998-2000	Supervision of B, C- and D-students, University of Umeå, Sweden.
1997	University of Iceland, Women's Studies
1994-1995	Teaching assistant in two courses on social history, University of Iceland, History Department
1987-1989	Public school teacher, Réttarholtsskóli, Reykjavík, Iceland

Publications related to proposal

Ó. Garðarsdóttir. 2004. "Þurfamenn í manntalinu 1703" ("Paupers in the 1703 Iceland census"). A seminar organized by the National Archives of Iceland and Statistics Iceland to commemorate the 300th anniversary of the first Iceland census 1703. 15 November 2003.

E. Roberts, S. Ruggles, L. Dillon, Ó. Garðarsdóttir, J. Oldervoll, G. Thorvaldsen and M. Woollard. 2002. "The North Atlantic Population Project: An Overview." *Historical Methods.* 36(2): 80-88.

Ó. Garðarsdóttir. 2000. "The implications of illegitimacy in late nineteenth-century Iceland. The relationship between infant mortality and the household position of mothers giving birth to illegitimate children." *Continuity and Change.* 15(3): 435-461.

Ó. Garðarsdóttir and G.Á. Gunnlaugsson. 1996. "Transition into widowhood: life course perspective on the household position of Icelandic widows at the turn of the century." *Continuity and Change.* 11(3): 435-459.

Ó. Garðarsdóttir and G.Á. Gunnlaugsson. 1995. "Availability of Offspring and the Household Position of Elderly Women: Iceland 190.1" *Journal of Family History.* 20(2): 159-179

Other significant publications

Ó. Garðarsdóttir and L. Guttormsson. 2004 (forthcoming). "A successful intervention. The fight against infant mortality from neonatal tetanus in the island of Vestmannaeyjar (Iceland) during the first half of the nineteenth century."

Ó. Garðarsdóttir. 2004. "Barnmorskor, mammor och amningens betydelse för hälsa och spädbarns överlevnad i Island 1850-1930." (In Swedish) "Midwives, mothers and the importance of breastfeeding for health and survival of infants in Iceland 1850-1930" in A. Brändström, S.

Edvinsson and P. Sköld. (eds). *Befolkningshistoriska perspektiv. Festskrift till Lars-Göran Tedebrand*. Umeå.

Ó. Garðarsdóttir. 2002. *Saving the child. Regional, cultural and social aspects of the infant mortality decline in Iceland, 1770-1920*. Umeå: Umeå University.

Ó. Garðarsdóttir. 1999. "Naming Practices and Kinship Networks in Early Nineteenth-Century Iceland." *History of the Family*. 4(3): 297-314.

Ó. Garðarsdóttir. 1998. "Tengsl þéttbýlismyndunar og Vesturheimsferða frá Íslandi. Lýðfræðilega sérkenni fólksflutninga frá Seyðisfirði" (In Icelandic) "Demographic characteristics of the emigration from Iceland to North-America." *Saga.* 36: 153-184.

Synergistic Activities

1. Participant in the NAPP project from its start in 2001.
2. Initiator of a committee to commemorate the 300th anniversary of the 1703 census in Iceland.
3. Responsible for publication of population statistics in Iceland as head of population statistics at Statistics Iceland, Reykjavík.
4. Participant in research project "Frihet og likhet for velferdsstatenes barn? Nordiske barndommer 1900 – 2000" (Childhoods in the Nordic countries during the 20th century). Sponsored by NOS-H.

Collaborators within the last 48 months

Astri Andersen	University of Bergen	Catherine A. Fitch	University of Minnesota
Anders Brändström	University of Umeå	Chad Gaffield	University of Ottawa
Nanna Floor Clausen	DDA Copenhagen	Todd Gardner	U.S. Census Bureau
Mats Danielsson	University of Umeå	Eiríkur Guðmundsson	National Archives of Iceland
Marie Digoix	INED Paris		
Lisa Y. Dillon	University of Montreal	Loftur Guttormsson	University of Education
Sören Edvinsson	University of Umeå		
Marianne J. Erikstad	University of Tromsø	Guðmundur Hálfdanarson	University of Iceland
Patrick Festy	INED Paris		
Monika Janfelt	University of Odense	Kari Pitkänen	University of Helsinki
Benedikt Jónsson	National Archives	Evan Roberts	University of Minnesota
Iceland		John Rogers	University of Uppsala
Marco van Leuwen	IISG Amsterdam	Steven Ruggles	University of Minnesota
Cecilia Lindgren	University of Linköping	Indgrid Söderlind	University of Linköping
Pirjo Markkola	University of Tampere	Gunnar Thorvaldsen	University of Tromsø
Jan Oldervoll	University of Bergen	Matthew Woollard	University of Essex

Graduate and Post Doctoral Advisors

Anders Brändström	University of Umeå

Thesis Advisor and Postgraduate-Scholar Sponsor

N/A

Biosketch
Jan Oldervoll

Associate professor, Department of History
University of Bergen
Sydnesplassen 7,
N-5007 Bergen
Norway

Phone: (+47) 5558-2322
Fax: (+47) 5558-8600
E-mail: oldervoll@hi.uib.no

Professional Preparation

University of Bergen	History	BA, 1968
University of Bergen	History	MA, 1970

Appointments

1994-97	President, The International Association for History and Computing
1992-	Associate professor, Department of History, University of Bergen
1989	Visiting professor, Max-Planck-Institut für Geschichte, Göttingen, Germany
1988-92	Associate professor, Department of Social Sciences, University of Tromsø
1988-91	President, The Nordic Branch of the Association for History and Computing
1978-88	Associate professor, Department of History, University of Bergen
1969-78	Research assistant, Department of History, University of Bergen

Publications Related to Proposal

J. Oldervoll. 1992. "CensSys - A standardization tool?" In J. Smets (ed). *Histoire et Informatique. Ve Congrès History and Computing* 4-7 Septembre à Montpellier, p. 209-213.

J. Oldervoll (ed). 1992. *Eden or Babylon? On Future Software for Highly Structured Historical Sources*. (Göttingen: Max-Planck-Institut für Geschichte in Kommission bei Scripta Mercaturae Verlag).

J. Oldervoll. 1992. "Wincens, a Census system for the nineties?" In J. Oldervoll (ed). *Eden or Babylon? On Future Software for Highly Structured Historical Sources*. (Halbgraue Reihe zur Historischen Fachinformatik. Serie A: Historische Quellenkunden Band 13): 37-52.

J. Oldervoll. 1987. "How to 'Internationalize' the 1801-Census of Norway or A Proposal for an International Standard for Census Type Data." In F. Hausmann, R. Härtel, I. H. Kropac, P.Becker (eds). *Datennetze für die Historischen Wissenschaften? Probleme und Möglichkeiten bei Standardisierung und Transfer Maschinenlesbarer Daten* (Graz: Leykam-Verlag): 185-193.

J. Oldervoll. 1985. "Automatic Record Linkage of 18th Century Nominal Records." In R. F. Allen (ed). *The International Conference on Data Bases in the Humanities and Social Sciences*. (Osprey, FL: Paradigm Press): 335-340.

Other Significant Publications

J. Oldervoll. 1991. "The Machine-Readable Description of Highly Structured Historical Documents: Censuses and Parish Registers." In D. I. Greenstein (ed). *Modelling Historical Data: Towards a Standard for Encoding and Exchanging Machine-Readable Texts*. Halbgraue Reihe zur Historischen Fachinformatik, Serie A: Historische Quellenkunden, Band 11: 169-178.

J. Oldervoll. 1989. "CENSYS - A System for Analyzing Census-Type Data." *Historical Social Research*. 14(3):17–22.

Synergistic Activities

1. Co-director of the Digital Archive of Norway (http://digitalarkivet.uib.no), in charge of data presentation part of the project (1998 – present). The Digital Project has as its goal to digitize and present the holdings of the National Archives on the web for public use. We are serving a wide audience and are one of the most popular sites in Norway, with almost 100,000 requests a day, more than 20% of which are coming from United States. On all projects I did all the programming as well as running the project.

2. Director of the Kark project (http://kark,uib.no) since 1992. Kark is developing web-based systems for computer assisted learning and is presently used by more 3000 students at more than half a dozen universities and colleges in Scandinavia. Part of the project has been a total restructuring of the undergraduate teaching at the History department.

3. Director of the 1801 Project (1969 to 1978). This project digitized, and coded the 1801 National Census of Norway. The project was a cooperation between The National Archive of Norway, the Statistical Bureau of Norway and University of Bergen.

Collaborators within the last 48 months

Lisa Dillon	Université de Montréal
Marianne Erikstad	University of Tromsø
Chad Gaffield	University of Ottawa
Ólöf Garðarsðottir	Statistics Iceland
Simon Renton	University of London, UK
Evan Roberts	University of Minnesota
Irena Rozman	University of Ljubliana, Slovenia
Steven Ruggles	University of Minnesota
Kevin Schürer	University of Essex
Arne Solli	University of Bergen
Svein Sture	University of Bergen
Manfred Thaller	University of Cologne, Germany
Gunnar Thorvaldsen	University of Tromsø
Frode Ulvund	University of Bergen
Matthew Woollard	University of Essex

Graduate and Postgraduate Advisors

Knut Mykland	University of Bergen

Thesis Advisor and Postgraduate-Scholar Sponsor

Herdis Kolle	University of Bergen
Bent Opheim	University of Bergen
Raivo Ruusalepp	University of Bergen
Arne Solli	University of Bergen
Frode Ulvund	University of Bergen

Biosketch
Evan Roberts

Minnesota Population Center
University of Minnesota, Twin Cities
537 Heller Hall, 271 19th Avenue South
Minneapolis, MN 55455-0406

Phone: (612) 624-5818
Fax: (612) 626-8375
E-mail: eroberts@pop.umn.edu

Professional Preparation

Victoria University of Wellington	History/Economics	B.A., 1995
Victoria University of Wellington	Mathematics/Statistics	B.Sc., 1998
Victoria University of Wellington	History/Economics	BA(Hons), 1999
University of Minnesota	History	M.A., 2003
University of Minnesota	History	Ph.D., 2006 (expected)

Appointments

2001-	Research Assistant, Minnesota Population Center, University of Minnesota
2002-2003	Teaching Assistant, History Department, University of Minnesota
1997-2000	Research Assistant, Health Services Research Centre, Victoria University of Wellington
1996-1997	Analyst, Sector Analysis division, New Zealand Ministry of Health

Publications Related to Proposal

E. Roberts, S. Ruggles, L. Dillon, Ó. Garðarsdóttir, J. Oldervoll, G. Thorvaldsen and M. Woollard. 2002. "The North Atlantic Population Project: An Overview." *Historical Methods.* 36(2): 80-88.

E. Roberts, M. Woollard, C. Ronnander, L. Dillon, and G. Thorvaldsen. 2002. "Occupational Classification in the the North Atlantic Population Project." *Historical Methods.* 36(2): 89-96.

Other Significant Publications

E. Roberts, J. Cumming and K. Nelson. (forthcoming). "A Review of Economic Evaluations of Community Mental Health Care." *Medical Care Research and Review*.

E. Roberts. 2003. "'Don't Sell Things, Sell Effects': Overseas influences in New Zealand department stores, 1909-1956." *Business History Review.* 77 (Summer): 265-289.

E. Roberts. 2002. "Gender in Store: The Politics of Salespeople's Working Hours, and Retail Unions in New Zealand and the United States, 1930-1960." *Labour History.* (83):107-130.

E. Roberts and P. Norris. 2001. "Regional variation in anti-depressant dispensings in New Zealand: 1993-1997." *New Zealand Medical Journal.* 114(1125): 27-29.

E. Roberts and P. Norris. 2001. "Growth and change in the prescribing of anti-depressants in New Zealand: 1993-1997." *New Zealand Medical Journal.* 114(1125): 25-26.

Synergistic Activities

1. Coordinator of North Atlantic Population Project from 2001– present.
2. Occupational coder for Historical Census Projects, Minnesota Population Center, 2001 – present.
3. Secretary, International Microdata Access Group, October 2002 – present.
4. Organizer, with Lisa Y. Dillon, of the second IMAG international workshop titled "Longitudinal and Cross-Sectional Historical Data: Intersections and Opportunities." The conference was held at the

Université de Montréal (November 10-11, 2003) and featured scholars from Canada, the U.S., the Netherlands, Sweden, Belgium and Switzerland.

Collaborators within the last 48 months

Jacqueline Cumming	Victoria University of Wellington
Lisa Y. Dillon	Université de Montréal
Marianne Erikstad	University of Tromsø
Chad Gaffield	University of Ottawa
Ólöf Garðarsdóttir	Statistics Iceland
Kris Inwood	Guelph University
Katherine Nelson	Victoria University of Wellington
Jan Oldervoll	University of Bergen
Chad Ronnander	University of Minnesota
Steven Ruggles	University of Minnesota
Kevin Schurer	UK Data Archive
Gunnar Thorvaldsen	University of Tromsø
Hannelore Vandebroek	Catholic University of Leuven
Matthew Woollard	University of Essex

Graduate and Post Doctoral Advisors

Steven Ruggles	University of Minnesota

Thesis Advisor and Postgraduate-Scholar Sponsor

N/A

Biosketch
Kevin Schürer

Professor, Department of History
University of Essex,
Wivenhoe Park,
Colcester, Essex, CO4 3SQ. UK

Phone: (+44-1206) 872-000
Fax: (+44-1206) 872-003
Email: schurer@essex.ac.uk

Professional Preparation

University of North London	History and Geography	B.A., 1979
University of Cambridge	History	M.A., 1986
University of London	History	Ph.D., 1984

Appointments

2000- University Professor, Department of History, University of Essex and Director, UK Data Archive
1993-1996 Assistant Director, UK Data Archive
1982-1997 Senior Research Associate, Cambridge Group for the History of Population and Social Structure, University of Cambridge

Professional Engagement

2002-present Elected member, Society of Archivists
2002-present Elected Academician, Academy for Social Sciences
2002-present Elected Fellow, Royal Statistical Society
1999-present Elected Fellow, Royal Geographical Society

Publications Related to Proposal

K. Schürer. 2003. "Leaving home in England and Wales 1850-1920." in F. van Poppel, M. Oris and J. Lee (eds). *The Road to Independence. Leaving Home in Eastern and Western Societies, 16th-20th Denturies* (Bern-Bruxelles: Peter Lang).

K. Schürer. 2003. "Household and family in past time – further explored." *Continuity and Change.* 18(1): 9-21.

K. Schürer and L. Dillon. 2003. "What's in a name? Victorias in late nineteenth-century Great Britain and Canada." *Local Population Studies*. 70: 57-62.

E. Garrett, A. Reid, K. Schürer and S. Szreter. 2001. *Changing Family Size in England and Wales. Place, Class and Demography, 1891-1911*. Cambridge: Cambridge University Press.

K. Schürer. 2000. *Leaving home and the Labour Market: The Experience of England and Wales, 1850-1920.* Swindon.

Other Significant Publications

K. Schürer. 1999. "History on the Internet and in the Global Village." in S. Kolb and P. Rösgen (eds). *The Culture of European History in the 21st Century*. (Bonn : Haus der Geschichte der Bundesrepublik Deutschland).

K. Schürer. 1997. "Historical Demography, Social Structure and the Computer." in P. Denley and D. Hopkin, *History and Computing* (Manchester: Manchester University Press).

K. Schürer and D.R. Mills (eds). 1996. *Local Communities in the Victorian Census Enumerators' Books.* Oxford: Leopard's Head Press.

K. Schürer. 1991. "The 1891 census and local population studies." *Local Population Studies*. 47: 16-29.

K. Schürer. 1991. "The role of the family in the process of migration." In C. G. Pooley and I. D. Whyte (eds). *Migrants, Emigrants and Immigrants: A Social History of Migration* (London: Routledge).

Synergistic Activities

1. Director, U.K. Data Archive and Economic and Social Data Service.
2. Principal Investigator, 1881 British census project.
3. Principal investigator, "The Victorian Panel Study: a feasibility project." Economic and Social Research Council (£100,000, 2004 – present).
4. Principal investigator, "The future of work: an historical perspective." Economic and Social Research Council (£218,000, 1998 – 2002).
5. Principal investigator, "The Nineteenth Century Censuses Collection: a computerised resource for research and teaching." Leverhulme Trust (£145,000, 1998 – 2002).

Collaborators. within the last 48 months

Lisa Y. Dillon	Université de Montréal	Steven Ruggles	University of Minnesota
Eilidh Garrett	Cambridge University	Richard Smith	Cambridge University
Chad Gaffield	University of Ottawa	Humphrey Southall	University of Portsmonth
Ólöf Garðarsdóttir	Statistics Iceland		
Myron P. Gutmann	University of Michigan	James A. Sweet	University of Wisconsin
Robert McCaa	University of Minnesota	Simon Szreter	Cambridge University
Jan Oldervoll	University of Bergen	Gunnar Thorvaldsen	University of Tromsø
Alice Reid	Cambridge University	Matthew Woollard	University of Essex
Evan Roberts	University of Minnesota		

Graduate and Post Doctoral Advisors

Peter Laslett	University of Cambridge (deceased)
Philip Ogden	University of London
Eleanor Vollans	University of London (retired)

Thesis Advisor and Postgraduate-Scholar Sponsor

R. Cassidy	University of Essex
B. Cook	University of Essex
S. Gekis	University of Essex
C. Jones	University of Essex
S. Sovic	University of Essex
D. Townsend	University of Essex

Biosketch
Gunnar Thorvaldsen

Norwegian Historical Data Centre
University of Tromsø
N 9014 Tromsø
Norway

Office: (+47) 7764-4179
Fax: (+47) 7764-4182
Email: gunnarth@sv.uit.no
Web: http://www.rhd.uit.no

,

Professional Preparation

University of Tromsø	History	Dr. philos	1995
University of Oslo	History	Cand. philol	1978
University of Oslo	Sociology, English, Pedagogy	Cand. mag	1974

Appointments

2000-	Professor and Director, Norwegian Historical Data Centre, University of Tromsø
1991-2000	Associate professor and Director, Norwegian Historical Data Centre, University of Tromsø
1987-1991	Senior archivist, Norwegian National Archives, Oslo
1986-1987	Data consultant, the newspaper *Aftenposten*
1977-1986	Director and data consultant, Norwegian Historical Data Centre, University of Tromsø

Publications Related to Proposal (in English)

G. Thorvaldsen and S. Sogner. 2002. "Surnames as Proxies for Place of Origin in the 1801 census for Norway." In J Carling (ed). *Report from the 2001 Nordic Demographic Symposium*. Scandinavian Population Studies XIII.

P. Kelly Hall, R. McCaa and G. Thorvaldsen (eds). 2000. *A Handbook of International Historical Microdata for Population Research.* Minneapolis: Minnesota Population Center.

G. Thorvaldsen. 2000. "The Historical Census as a Research Tool." *Report from the Nordsaga Conference.* University of Tromsø.

G. Thorvaldsen. 1998. "Historical Databases in Scandinavia." *The History of the Family. An International Quarterly,* 3(3): 371-383.

G. Thorvaldsen. 1995. "The Encoding of Highly Structured Historical Sources." *Computers and the Humanities,* 28(4-5): 301-305.

Other Significant Publications (in English)

G. Thorvaldsen. 2002. "Coastal Women and their Work Roles." In Sandvik, Telste and Thorvaldsen (eds*) Pathways of the Past. Essays in Honour of Sølvi Sogner on her 70th Anniversary*.

G. Thorvaldsen. 2001. "Northern, Eastern Or Urban Penalty? Aspects of Late 19th Mortality in the North of Norway." In Lars-Göran Tedebrand & Peter Sköld (eds). *Report from the 1998 Nordic Demographic Symposium*.

R. Miller and G. Thorvaldsen. 1997. "Beyond Record Linkage: Longitudinal Analysis of Turn-of-the-century Interurban Swedish Migrants." *History and Computing*. 9(1-3): 106-121.

G. Thorvaldsen. 1997. "Marriage and Names among Immigrants to Minnesota." *Electronic Journal of the the American Association for History and Computing* http://mcel.pacificu.edu/history/JAHC/Thorvaldsen/ThorIndex.HTML.

G. Thorvaldsen. 1995. Doctoral dissertation: *Migration in the province of Troms 1865-1900. A Study based on the Censuses.* Dr. philos avhandling: *Migrasjon i Troms i annen halvdel av 1800-tallet. En kvantitativ analyse av folketellingene 1865, 1875 og 1900*. With summary in English.

Synergistic Activities

1. Coordinator of project with University of Dar es Salaam "Time-use by gender studied with survey and census data in contemporary Tanzania" (2003-2007).
2. Editing two publications on the history of the Norwegian census and building a related website with Statistics Norway (http://www.rhd.uit.no/census.htm) (2000-present).
3. Chairman of the Norwegian Demographic Society (2001- present).
4. Board member, The International Association for History & Computing (1994-2003).
5. Principal investigator, "Geographical Mobility in the Province of Troms," funded by the Norwegian Research Council. ($100.000, 1990-1994).

Collaborators within the last 48 months

Gulbrand Alhaugh	University of Tromsø
Trygve Andersen	University of Tromsø
Lisa Dillon	Université de Montréal
Sören Edvinsson	Umeå University
Marianne Erikstad	University of Tromsø
Ólöf Garðarsðottir	Statistics Iceland
Elena Glavatskaia	Ural State University
William Hubbard	University of Bergen
Roger Miller	University of Minnesota
Lars Nygaard	National Archives, Oslo
Kari Pitkänen	University of Helsinki
Evan Roberts	University of Minnesota
Chad Ronnander	University of Minnesota
Teocratias Rugaimakumu	University of dar es Salaam
Steve Ruggles	University of Minnesota
Sølvi Sogner	University of Oslo
Espen Søbye	Statistics Norway

Graduate and Postgraduate Advisor

Sivert Langholm	University of Oslo

Thesis Advisor and Postgraduate-Scholar Sponsor

Lisbeth B Johansen	University of Tromsø
Hilde L Granly	University of Tromsø

Biosketch
Matthew Woollard

AHDS History, UK Data Archive
University of Essex
Colchester, Essex
UK CO4 3SQ

Phone: (+44-1206) 873-704
Fax: (+44-1206) 872-003
E-mail: matthew@essex.ac.uk

Professional Preparation

Goldsmiths' College, University of London,	History	B.A., 1990
Institute of Historical Research,	Computer applications for historians	M.A., 1991
University of London, University of Essex.	History	Ph.D in progress

Appointments

2002–	Head of Service, Arts and Humanities Data Service History, University of Essex. Project Director, Online Historical Population Reports Project.
2001–2002	Senior Research Officer/Research Associate, University of Essex. incl. North Atlantic Population Project.
1998–2001	Senior Research Officer, ESRC Future of Work project, University of Essex.
1996–1998	Research Officer, Nineteenth-century censuses project, University of Essex.
1993–1996	Research Associate, Bristol Historical Databases project, University of the West of England, Bristol.
1992–1993	Research Assistant, English Version of kleio Project, University of Southampton and Queen Mary and Westfield College, University of London.

Publications Related to Proposal

E. Roberts, S. Ruggles, L. Dillon, Ó. Garðarsdóttir, J. Oldervoll, G. Thorvaldsen and M. Woollard. 2002. "The North Atlantic Population Project: An Overview." *Historical Methods. 36(2)*: 80-88.

E. Roberts, M. Woollard, C. Ronnander, L. Dillon, and G. Thorvaldsen. 2002. "Occupational Classification in the the North Atlantic Population Project." *Historical Methods. 36(2)*: 89-96.

M. Woollard. 2000. "Microdata from the 1851 and 1881 censuses" in P. Kelly Hall, R. McCaa and G. Thorvaldsen (eds). *Handbook of international historical microdata for population research*. (Minneapolis: Minnesota Population Center): 107–121.

M. Woollard. 2001. "The 1901 census: an introduction." *Local Population Studies*. 67: 26–43.

M. Woollard. 1996. "Record linkage revisited." in M. Woollard and P. Denley (eds). *The sorcerers apprentice: kleio case studies* (St. Katharinen: Max-Planck-Institut für Geschichte in Kommission bei Scripta Mercaturae Verlag): 151–98.

Other Significant Publications

M. Woollard. 2004. "The classification of multiple occupational titles in the 1881 census of England and Wales." *Local Population Studies.* (72): 34–49.

M. Woollard. 2002. "The employment and retirement of older men, 1851–1881: further evidence from the census." *Continuity and Change*. 17(3): 437–463.

M. Woollard. 1999. "What is history and computing? An introduction to a problem." *History and Computing*. 11(1): 1–8.

M. Woollard & P. Denley. 1993. *Source-oriented data processing for historians: a tutorial for kleio*. St. Katharinen: Max-Planck-Institut für Geschichte in Kommission bei Scripta Mercaturae Verlag.

Synergistic Activities

1. Project Director and principal investigator of Online Historical Population Reports Project, a project to digitize and make available on the web all British census reports, (1801-1931) and all published official historical demographic reports (http://www.histpop.org.uk). Joint Information Systems Committee, £912,000 (2004- 2007).

Collaborators within the last 48 months

Lisa Y. Dillon	University of Montreal
Chad Gaffield	University of Ottawa
Ólöf Garðarsdóttir	Statistics Iceland
Eddy Higgs	University of Essex
Kris Inwood	University of Guelph
Andrew Miles	University of Birmingham
Jan Oldervoll	University of Bergen
Evan Roberts	University of Minnesota
Chad Ronnander	University of Minnesota
Steve Ruggles	University of Minnesota
Kevin Schürer	University of Essex
Gunnar Thorvaldsen	University of Tromsø
Marco van Leeuwen	IISG
Richard Wall	University of Essex

Graduate advisor

Eddy Higgs	University of Essex