MPC
Minnesota Population Center
UNIVERSITY OF MINNESOTA

Building a National Longitudinal Research Infrastructure

Steven Ruggles†
University of Minnesota

Catherine Fitch
University of Minnesota

Matthew Sobek
University of Minnesota

December 2017

Working Paper No. 2017-2
https://doi.org/10.18128/MPC2017-2

†Correspondence should be directed to:
Steven Ruggles
University of Minnesota, 50 Willey Hall, 225 19th Ave S., Minneapolis, MN 55455
e-mail: ruggles@umn.edu, phone: 612-624-5818, fax:612-626-8375

NLRI collaboration members: J. Trent Alexander, University of Michigan, Martha Bailey, University of Michigan, Joseph Ferrie, Northwestern University, Katie Genadek, U.S. Census Bureau, David Grusky, Stanford University, J. David Hacker, University of Minnesota, Michael Hout, New York University, Amy B. O'Hara, U.S. Census Bureau, Evan Roberts, University of Minnesota, Seth Sanders, Duke University, and John Robert Warren, University of Minnesota

This paper describes a new initiative to create and disseminate longitudinal data infrastructure for the United States based on the entire population enumerated between 1850 and 2020. The National Longitudinal Research Infrastructure (NLRI) aims to produce a foundational reference collection for demographic and health research. The availability of a massive collection of life histories of the U.S. population over 170 years will open new avenues for social and behavioral research, education, and policy-making. By disseminating the infrastructure to the broadest possible audience, the project will enhance scientific and public understanding of critical policy-related issues.

We are developing the infrastructure through three closely interconnected research projects: (1) the Census Longitudinal Infrastructure Project (CLIP); (2) the American Opportunity Study (AOS); and (3) the Multi-Generational Longitudinal Panel (IPUMS-MLP). The paragraphs that follow briefly describe the origins of the project and our preliminary studies. We then explain how NLRI will overcome critical barriers and transform research on the effects of public policies, social institutions, and health care on the health, well-being, and functioning of people over the life course and in their later years.

Background

NLRI builds on the work of the IPUMS project at the Minnesota Population Center (MPC), which pioneered novel methods for large-scale data integration and dissemination. IPUMS demonstrated that a long series of large integrated U.S. census microdata samples provides powerful tools for analyzing demographic and economic processes. IPUMS has become one of the most intensively-used data resources in the world. Over the past two decades, IPUMS has been used by over 100,000 researchers. These investigators currently download about 2.6 terabytes of

IPUMS data per week, which they use to produce some 1,400 papers each year across a broad range of disciplines (Google Scholar 2016).

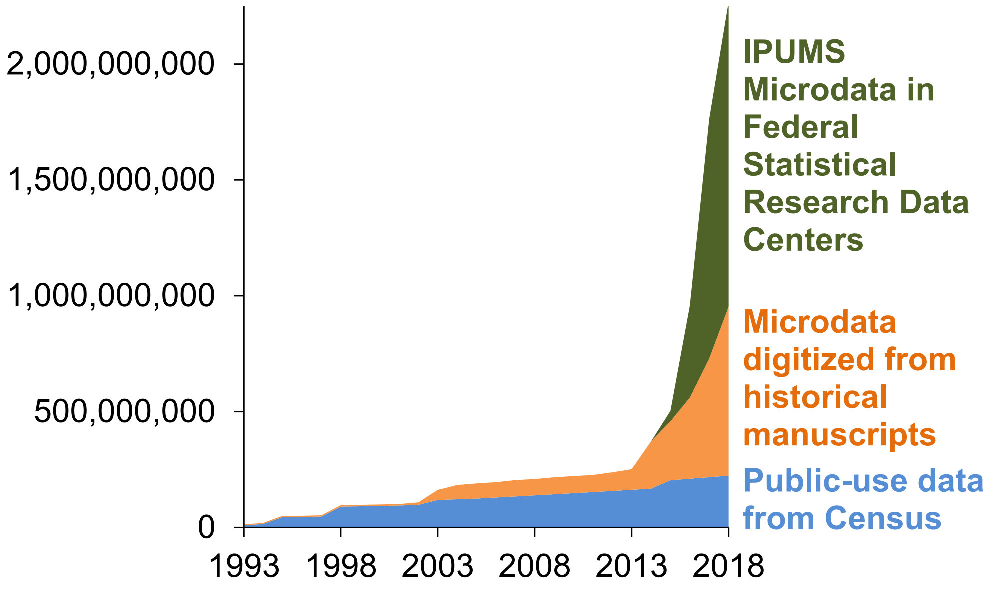
The signature activity of IPUMS is to harmonize data across time and place, so the same codes have the same meaning for all datasets in the collection. From its beginnings as an integrated database of public-use U.S. census samples, IPUMS has grown into a suite of projects that harmonize census and survey microdata from the United States and around the world. The NLRI initiative described here is a direct outgrowth of the original U.S. census project.

The U.S. IPUMS data collection is growing explosively thanks to two major new initiatives. Under the "Big Microdata" project, Ancestry.com donated complete-count U.S. census microdata spanning the period 1790-1940 to the scientific community. We estimate that these data would have cost over a half-billion dollars to replicate using conventional methods. Ancestry originally entered only variables of particular interest for genealogy. Now, with the support of Ancestry, the National Institutes of Health, and the National Science Foundation, we have enhanced the files to incorporate virtually all the variables originally enumerated and we are now converting the data into a format suitable for use by the scientific community. This work is well underway and is scheduled to be complete by 2019 (Ruggles 2014).

Simultaneously, under the Census Bureau's "National Historical Census Files" project, we are converting all internal-use U.S. Census Bureau microdata from 1960 to the present into standardized IPUMS format. As part of this project, we restored missing long-form data from the 1960 census by recovering data from microfilm using optical mark recognition (Ruggles et al. 2011). The IPUMS-format internal microdata—including the American Community Surveys as well as the Decennial Censuses—will become available in the Federal Statistical Research Data Centers (FSRDCs) in 2018.

Figure 1 shows the number of person-records of U.S. IPUMS data from the first data release in 1993 through 2018. At this writing, the total is over a billion records; three years hence, the total will exceed two billion.

Figure 1. Integrated U.S. microdata available for research 1993-2018 (number of person records)



Despite its high impact, IPUMS suffers from a profound limitation: each of the censuses is an independent cross-section. IPUMS is invaluable for studying period and cohort change, but the existing database cannot address life-course change. This handicap precludes using IPUMS to study the impact of early life condition on later outcomes. Moreover, the lack of longitudinal information sharply limits the potential for causal inference. NLRI is designed to overcome these limitations. The complete machine-readable census enumerations provide the opportunity for a national longitudinal panel that traces individuals over their lives and families over multiple generations.

To transform the massive series of census microdata files into a longitudinal data structure, we have assembled a team of the world's leading experts in automatic record linkage of censuses and administrative records. This project leverages data resources and linking capabilities of the Census Bureau, the data infrastructure expertise of the Minnesota Population Center, and an unparalleled team of experts from across the United States.

For the past four decades the U.S. Census Bureau has been at the forefront of innovation in automatic record linkage (Jaro 1972; Winkler 1989, 1999). Under the leadership of Amy O'Hara, the Census Bureau's Center for Administrative Records Research and Applications (CARRA) has developed unprecedented capabilities for large-scale matching of restricted data (Johnson et al. 2015; Massey 2014a, 2014b; Massey and O'Hara 2015). Academic researchers on our team are leading developers of technology for automatic linkage of historical census records. Joseph Ferrie (1996) pioneered large-scale linkage of historical censuses, and the technology has been improved through application of machine-learning technology by Peter Christen, Steven Ruggles, and Ronald Goeken (Christen 2012; Ruggles, 2006, 2011; Goeken et al. 2011).

We have conducted extensive preliminary studies that demonstrate the project's feasibility. CARRA's Census Longitudinal Infrastructure Project (CLIP), with support from the Census Bureau, MPC, and an NIH Exploratory/ Developmental Grant Award, has developed the necessary strategies for constructing NLRI's framework by linking the 1940 census to administrative records and to the 2000 and 2010 censuses. The American Opportunity Study (AOS), with support from the National Research Council and the National Science Foundation, has demonstrated the feasibility of linking the 1990 census to other sources and has conducted preliminary research on technology needed to link the censuses of 1960 through 1980 (Grusky et al. 2015). Ongoing research at MPC and the University of Michigan is streamlining and refining machine-learning technology for historical record linkage, with innovations to improve efficiency and exploit high-performance computing capacity.

Needs and Opportunities

Unlike some other developed countries, the United States lacks a large-scale longitudinal data source covering the entire population, limiting the efficacy and depth of analyses of population aging and life-course health. NLRI will address this need, going far beyond the usual capabilities of register-based data resources: longitudinal data of this depth have never existed for any country. NLRI will consist of linked census, survey, and administrative records covering the entire U.S. population over the past century, together with software enabling construction of customized datasets tailored to specific research problems. NLRI will be invaluable for analyzing the impact of early life conditions on health and well-being in later life. The large scale of the resource will allow study of very small population subgroups, including the oldest-old.

Former Census Bureau Director Robert Groves (2011) drew an insightful distinction between “designed data” and “organic data.” Designed data, such as censuses and surveys, are created entirely to obtain information. Organic data are byproducts of transactions, including administrative records generated by Social Security, Medicare, the Internal Revenue Service, and the Armed Forces. Research on population aging currently relies primarily on designed data, despite the enormous potential of organic data to enrich our analyses.

Groves argued that “the biggest payoff will lie in new combinations of designed data and organic data, not in one type alone.” Used in isolation, organic data have profound limitations that limit their usefulness. They tend to be voluminous but shallow; they often are unrepresentative of the general population, and they frequently omit basic information about demographic behavior, economic status, education, work, and living conditions. NLRI will enrich some of the largest sources of organic data—including Social Security, Medicare, and military records—by linking them to designed census and survey data, thereby overcoming limitations of the organic data sources.

Linking individuals from childhood to old age and death through both designed and organic data allows study of aging as a process over the entire life course, not just over a few years. Indeed, NLRI will enable investigators to extend longitudinal analysis beyond individual life histories, to investigate and understand processes of change over multiple generations. In his recent presidential address to the Population Association of America, Robert Mare (2011) argued that "the study of intergenerational mobility and most population research are governed by a two-generation (parent-to-offspring) view of intergenerational influence, to the neglect of the effects of grandparents and other ancestors and nonresident contemporary kin." Mare called for the development of sources and methods that will allow for analysis of change over multiple generations. NLRI meets this need, allowing investigators to trace records back across multiple generations and making it possible for the first time to study the transmission of demographic characteristics and behavior across centuries.

NLRI will multiply the value of existing aging surveys. NLRI can add historical depth to the most important surveys relevant to aging research, such as the Health and Retirement Study (HRS), the Panel Study of Income Dynamics (PSID), the Wisconsin Longitudinal Study (WLS), the National Social Life, Health, and Aging Project (NSHAP), the National Health and Aging Trends Study (NHATS), the National Health Interview Survey (NHIS), and the American Community Survey (ACS). Individual researchers with identified data derived from clinical trials, patient records, or surveys will be able to bring those data into the secure Census Bureau environment where these sources can be linked with NLRI, providing rich demographic and administrative data for individuals and their families, often over multiple generations. The linked data for 1940 and earlier will be available without restriction to researchers everywhere. For confidentiality reasons, more recent data must be accessed through the Federal Statistical Research Data Centers.

To meet the challenges created by rapid population aging, researchers must have full and open access to the best possible information. This infrastructure will support large-scale studies of population aging and life-course transitions. NLRI will multiply the quality, quantity, accessibility, and interoperability of information about the changing American population, creating a resource of unprecedented power for understanding ongoing transformations of demographic behavior. Accordingly, NLRI will become a core component of the U.S. statistical infrastructure.

NLRI is not designed to answer a particular scientific question. Rather, it is general-purpose data infrastructure, designed to become a permanent data resource that will be used for decades to come and can be continuously expanded to incorporate the latest data sources as they become available. Thousands of innovative analyses will become feasible; here are a few representative examples:

- *Impact of exposure to water-borne lead before age 3 on Late Onset Alzheimer's Disease (LOAD)*. Prince (1998) has suggested that, because lead and LOAD both show their effects in the same areas of the brain, lead might be an environmental source of a predisposition toward LOAD. This conjecture has been borne out in animal experiments but has not been assessed in human populations due to lack of information on lead exposure under age three, the period of greatest sensitivity identified in the animal experiments. NLRI will provide information about lead exposure in early childhood for millions of Medicare recipients through the well-known relationship between water pH and its plumbosolvency (Ferrie et al. 2012).

- *The receipt of Mothers' Pension benefits in childhood and later-life welfare recipiency and labor force participation*. Aizer et al. (2016) show a strong causal link between the receipt of Mothers' Pensions (the pre-1930 forerunner of AFDC/TANF) and a range of improved outcomes for children, such as weight, income, and longevity. NLRI will allow us to identify additional effects across the life course—on cognition, labor force participation, asset accumulation, and health at older ages—as well as the mechanisms through which these effects are produced.

- *Intergenerational transmission of health and well-being over multiple (3+) generations*. The relationship between outcomes for one generation (education, income, health) and those for earlier or later generations in the same family line has been examined in the U.S. almost exclusively in a two-generation (parent-child) context. This limitation is largely due to the paucity of data linking 3 or more generations (Pfeffer 2014). NLRI will include U.S. population censuses linked from 1850 through the present. This will permit both (1) assessing the influence of up to 6 previous generations on an individual's outcomes, potentially revealing the extent to which two-generation research has understated the persistence in outcomes across generations; and (2) analyzing trends over more than 150 years in two-generation mobility.

- *The impact of early-life cognitive capacity on later-life health and economic outcomes*. Research into the role of cognitive measures on later-life earnings, health, and longevity in the U.S. has not been possible for large populations because no records connect cognitive testing early in life to census or administrative records on later life outcomes. We have discovered IQ scores for 500,000 U.S. male enlistees in World War II, and will use NLRI to link them to subsequent censuses and earnings and health records, following these individuals from their

> mid-twenties to their deaths. Project Talent records on the IQ of 440,000 U.S. high school students tested in 1960 (Flanagan 1962) provide the capacity to do the same for cohorts born twenty years later. The World War II and Project Talent data also allow the analysis of intergenerational transmission of cognitive ability for a large representative sample of families.

By leveraging billions of dollars of federal investments in designed data collected over the course of two centuries, and combining them with organic transactional records from a variety of administrative sources, we can create a powerful new resource for research, education, and policy-making on health and aging. Researchers have never before had access to longitudinal data of this scale and scope anywhere in the world; this wealth of new data will spawn new methods of life-course analysis that can deepen understanding of the ongoing transformations of the American population and society.

Strategies for a National Longitudinal Data Infrastructure

NLRI will construct the world's largest longitudinal population database by linking key elements of the U.S. demographic and administrative records over 170 years. The project will merge billions of individual records from dozens of sources to construct millions of individual life histories.

NLRI's unprecedented breadth and depth is its central innovation. To achieve this conceptual innovation, however, requires dozens of technical innovations. Converting billions of records of raw data into a functional longitudinal data infrastructure requires new approaches to record linkage and big data processing. To cope with the scale, complexity, and heterogeneity of the data, we must engage the leading edge of computer and information science and develop innovative solutions through the entire data life cycle. We will build novel machine-learning data mining technology to construct national panels with repeated observations of individuals and families. We will also extend and improve our existing metadata systems and electronic dissemination tools to accommodate the new longitudinal files and to provide capacity for the massive scale of the collection.

NLRI consists of three closely interconnected projects that address distinct aspects of data infrastructure. Although each of the projects focuses on record linkage, their challenges differ. Because the information available for linking varies across the three NLRI components, each has its own unique linking strategies.

Project 1: Census Longitudinal Infrastructure Project (CLIP). Aims: Develop linked data from U.S. censuses, surveys, and administrative records spanning the period from 1940 to the present, and implement a big data access system at the Census Bureau.

This project will make it feasible for researchers to use linked datasets with information from multiple demographic and administrative sources through Federal Statistical Research Data Centers. The Research Data Centers make nonpublic federal data available to researchers at 24 locations around the United States. All record linkage pertaining to the period since 1940 is taking place within the Census Bureau's Center for Administrative Records Research and Applications. The framework of the infrastructure is a collection of Census Bureau demographic products with Protected Identification Keys (PIKs) that are associated uniquely with a specific individual and consistent across all data sources, allowing researchers to link individuals across surveys. The starting point is to attach PIKs to the complete 1940 census microdata now being finalized at the Minnesota Population Center. We are using the Social Security Administration's Numerical Identification System (Numident) to assign the PIKs. CLIP has already identified 72% of children age 0-9 who appear in the 1940 census, allowing investigators to understand the origins of late life outcomes among people who reached age 65 between 1995 and 2005.

The PIKs allow CLIP to link individuals from the 1940 census to the 2000 and 2010 censuses, to other Census Bureau demographic products, including the Annual Social and Economic Supplement of the Current Population Surveys and the American Community Surveys, and to a variety of administrative records such as the Medicare Enrollment Database and the Selective Service Registration system. The Department of Housing and Urban Development has recently funded a CLIP project to PIK all the American Housing Surveys back to 1973.

These linked datasets form the core framework for the integrated panel data. The core framework will be linkable to a wide range of additional records, including survey and administrative data from many sources. Already, nine substantive research projects using the CLIP infrastructure are underway.

Project 2: American Opportunity Study. Aims: Extend the CLIP Project to the 1990 census, and develop new technology that will allow incorporation of microdata from the 1960, 1970, and 1980 censuses.

The American Opportunity Study (AOS) is designed to assess change in social and economic mobility by linking the internal census microdata for the 1960 through 1990 period (Johnson, Massey, and O’Hara 2015). Unlike the 2000 and 2010 censuses, these earlier censuses did not digitize names at the time they were created. This greatly complicates record linkage, since there is no simple way to assign PIKs to the records. Accordingly, the project is pursuing two strategies to link these censuses. We have discovered a 1990 census file that contains street addresses, and these addresses can be matched to administrative records that provide names. In addition, we have uncovered a file containing names for census households in multifamily dwellings and rural areas. These two sources will allow us to PIK the overwhelming majority of adults in 1990. To fill in the rest of the population in 1990, we are developing cutting-edge natural handwriting recognition software that allows us to automate the digitization of names from scanned images of the original census enumeration forms.

We anticipate that automatic handwriting recognition technology, in combination with geocoded IRS tax records, will be the key to cost-effective PIKs for the 1960-1980 data. We anticipate that full linkage of these censuses will require additional resources, but the current project will develop and test the technology that will make such record linkage feasible.

Project 3: Multi-Generational Longitudinal Panel. Aims: Research and implement strategies for linking individuals and families across censuses from 1940 back to 1850, and disseminate these data freely to the public through IPUMS and through CLIP within the RDCs.

This project aims to extend CLIP backwards in time by tracing individuals and families across multiple generations. We will capitalize on an NIH-sponsored initiative to create census microdata covering the entire U.S. population from 1850 to 1940. Individuals and families can be traced backwards from 1940, allowing us to observe individual change over the entire life course and family change across multiple generations. This will be the largest population record linkage effort yet undertaken. To maximize success we must conduct new research to advance record linkage technology. Early versions of linked historical microdata have already shifted our perception of life-course change in the past. These linked historical census samples have revealed that occupational mobility was far higher in the 19th century than it is today, migration was much more frequent, and the formation of intergenerational families was most common among the rich (Long and Ferrie 2013; Ruggles 2011). The next generation of linked microdata will be far more powerful, offering 1000 times the number of records, more reliable links, and coverage across entire lives and multiple generations, allowing multilevel analysis of the demographic and economic context of mobility and family transitions. The resulting dataset will be a publicly-accessible resource and will be freely available through a web-based dissemination system. In addition, these linked historical microdata will be available within FSRDCs, where they can be joined with CLIP using 1940 as a crosswalk to enable analyses spanning the past 170 years.

Summary

Although the goals of each NLRI project are similar, the strategies differ substantially. The biggest challenge in Project 1 is producing PIKs for the 1940 census; Project 2 will PIK the 1990 census, and develop strategies to PIK the 1960 through 1980 censuses; and Project 3 aims to link the 1850 through 1940 censuses. These three groups of censuses have different information available for record linkage, and each must accordingly use different strategies. Despite the methodological differences, the three projects share a great deal in common, and there are numerous opportunities for cross-fertilization. Because all projects are using a shared framework, improvements and discoveries in any one of the projects will benefit all three.

The preliminary research for each of the three projects was conducted independently, and they could have proceeded as independent projects while still providing great scientific benefit. Joining these projects together in a single coordinated effort, however, has provided three major benefits. First, each project can learn from the others. Second, we can adopt common standards, formats, and metadata to make all components of the infrastructure interoperable. Third, and most important, in combination the three projects will constitute something much bigger than the sum of the parts, since every additional linked dataset enriches the infrastructure as a whole.

References

Aizer A, Eli S, Ferrie JP, and Lleras-Muney A. 2016. The Long-Run Impact of Cash Transfers to Poor Families. *American Economic Review* 106: 935-971.

Apache Parquet. 2016. Retrieved 15 January 2016 from parquet.apache.org.

Armbrust, M., Xin, R. S., Lian, C., Huai, Y., Liu, D., Bradley, J. K., Meng, X., Kaftan, T., Franklin, M. J., Ghodsi, A. & Zaharia, M. 2015. Spark Sql: Relational Data Processing in Spark. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*: 1383–94.

Barreca A, Clay K, and Tarr J. 2014. Coal, Smoke, and Death: Bituminous Coal and American Home Heating. National Bureau of Economic Research Working Paper w19881.

Christen P. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection.* New York: Springer.

Ferrie J. 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* 34: 141-56.

Ferrie JP, Rolf K, Troesken W. 2012. Cognitive Disparities, Lead Plumbing, and Water Chemistry: Prior Exposure to Water-Borne Lead and Intelligence Test Scores among World War Two U.S. Army Enlistees. *Economics and Human Biology* 10: 98–111.

Flanagan JC. 1962. *The Project Talent Data Bank: A Handbook*. Palo Alto, CA: American Institutes for Research in the Behavioral Sciences.

Goeken R, Huynh L, Lynch TA & Vick R. 2011. New Methods of Census Record Linking. *Historical Methods* 44: 7-14.

Google Scholar 2016. Search for IPUMS publications in 2015, Accessed 4/9/2016 https://scholar.google.com/scholar?q=IPUMS+OR+%22Integrated+Public+Use%22&hl=en&as_sdt=0%2C24&as_ylo=2015&as_yhi=2015

Groves R. 2011. Three Eras of Survey Research. *Public Opinion Quarterly* 75: 861-871.

Grusky DB, Smeeding TM & Snipp CM. 2015. A New Infrastructure for Monitoring Social Mobility in the United States. *Annals of the American Academy of Political and Social Science* 657: 63-82.

Jaro MA. 1972. UNIMATCH—A Computer System for Generalized Record Linkage under Conditions of Uncertainty. *Spring Joint Computer Conference, AFIPSLConference Proceedings* 40: 523–530

Johnson DS, Massey C, & O'Hara A. 2015. The Opportunities and Challenges of Using Administrative Data Linkages to Evaluate Mobility. *Annals of the American Academy of Political and Social Science* 657: 247-264.

Long J & Ferrie JP. 2013. Intergenerational Occupational Mobility in Great Britain and the United States since 1850. *American Economic Review* 103: 1109-1137.

Mare RD. 2011. A Multigenerational View of Inequality. *Demography* 48: 1-23. PMCID: PMC3059821.

Melnik S, Gubarev A, Long JJ, Romer G, Shivakumar S, Tolton M & Vassilakis T. 2011. Dremel: Interactive Analysis of Web-Wcale Datasets. *Communications of the ACM* 54: 114–23.

Massey CG. 2014a. Creating Linked Historical Data: An Assessment of the Census Bureau's Ability to Assign Protected Identification Keys to the 1960 census. CARRA Working Paper 2014-12.

Massey CG. 2014b. Playing with Matches: An Assessment of Match Accuracy in Linked Historical Data. CARRA Working Paper 2014-XX.

Massey CG & O'Hara A. 2014. Person Matching in Historical Files using the Census Bureau's Person Validation System. CARRA Working Paper 2014-11.

Pfeffer FT. 2014. Multigenerational Approached to Social Mobility: A Multifaceted Research Agenda. *Research in Social Stratification and Mobility* 35: 1-12.

Prince M. 1998. Is Chronic Low-Level Lead Exposure in Early Life an Etiologic Factor in Alzheimer's Disease? *Epidemiology* 9: 618-621.

Ruggles S. 2006. Linking Historical Censuses: A New Approach. *History and Computing* 14: 213-24.

Ruggles S. 2011. Intergenerational Coresidence and Family Transitions in the United States. *Journal of Marriage and Family* 73: 136–148.

Ruggles S. 2014. Big Microdata for Population Research. *Demography* 51: 287-297.

Ruggles S, Schroeder M, Rivers N, Alexander JT, & Gardner TK. 2011. Frozen Film and FOSDIC Forms: Restoring the 1960 census of Population. *Historical Methods* 44: 69-78

Winkler WE. 1989. Frequency-based Matching in the Fellegi-Sunter Model of Record Linkage," *ASA 1989 Proc. of the Section on Survey Research Methods*: 788-793.

Winkler WE. 1999. The State of Record Linkage and Current Research Problems. Technical Report Statistical Research Report Series RR99/04, US Bureau of the Census, Washington, D.C., 1999.