Building a National Longitudinal Research Infrastructure

Steven Ruggles, University of Minnesota
Catherine A. Fitch, University of Minnesota

Extended Abstract

This presentation will describe 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) will 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. The data represent a permanent and substantial addition to the nation's statistical infrastructure and will have far-reaching implications for research across the social and behavioral sciences. 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 two closely interconnected research projects: (1) the Census Longitudinal Infrastructure Project (CLIP); and (2) the Multi-Generational Longitudinal Panel. 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 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 almost 200,000 researchers. These investigators currently download about 5 terabytes of IPUMS data per week which they use to produce some 2,000 papers each year across a broad range of disciplines (Google Scholar 2017).

The IPUMS data collection is growing explosively thanks to two major new initiatives. Under the "Big Microdata" project, Ancestry.com and FamilySearch donated complete-count census microdata spanning the period 1790-1940 to the scientific community. With the support of Ancestry.com, NIA, NICHD, and NSF, we are now enhancing the files to incorporate virtually all the variables originally enumerated and 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 the end of 2018 (Ruggles 2014).

Simultaneously, under the Census Bureau's "National Historical Census Files" project, we are converting all internal 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 in 2017.

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

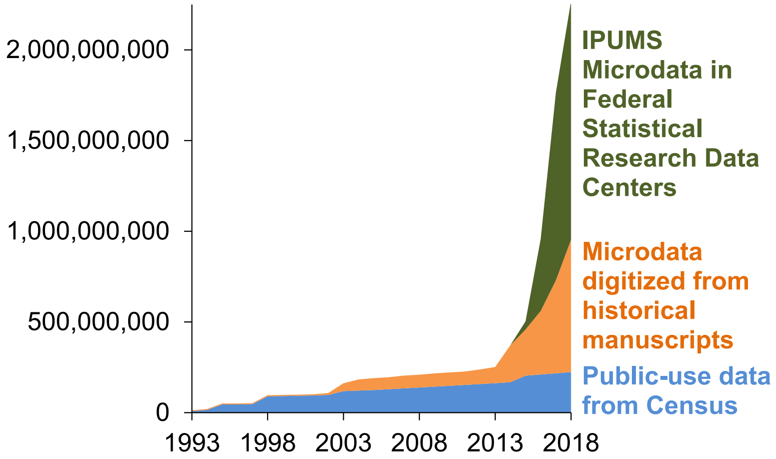

Figure 1 shows the number of person-records of IPUMS data from the first data release in 1993 through 2018. At this writing, the total is over a billion records; ultimately, the total will exceed two billion.

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.

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 the health and well-being in later life. The very 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 insolation, 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.

Approach

NLRI consists of two closely interconnected projects that will work on distinct aspects of NLRI infrastructure. Although each of the projects focuses on record linkage, the challenges of each project differ.

Project 1: Census Longitudinal Infrastructure Project (CLIP). Aims: Develop linked data from 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 (FSRDCs). Formerly called Census Bureau Research Data Centers, FSRDCs make nonpublic federal data available to researchers at 30 locations around the country. All record linkage pertaining to the period since 1940 will take place within the Census Bureau's Center for Administrative Records Research and Applications, and only de-identified data is made available to approve researchers working the secure FSRDCs. The key starting point is to link respondents to recent censuses to the complete 1940 census microdata now being finalized at the Minnesota Population Center with funding from NICHD, NIA, and NSF. A pilot project is already in process using a preliminary version of the 1940 data. 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 other sources.

Project 2: 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.

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.

Biographical Statement

Steven Ruggles is Regents Professor of History and Population Studies and Director of the Institute for Social Research and Data Innovation at the University of Minnesota. He is best-known as the inventor of IPUMS, the world's largest population database. IPUMS provides information about two billion people residing in 107 countries between 1703 and 2017, including every respondent to the surviving U.S. censuses of 1790 to 1940. Ruggles has published extensively on historical demography, focusing especially on long-run changes in multigenerational families, single parenthood, divorce, and marriage, and on methods for population history. He has received the Sharlin Award (SSHA), the Goode Award (ASA), the Lapham Award (PAA), and the Miller Award (ICPSR). In 1995, he was named "King of Quant" by Wired Magazine; in 2014, he was named "Wonkblog-Certified Data Wizard" by the Washington Post Wonkblog; and in 2009 he was honored with the coveted "Platinum Medallion" from Delta Airlines. He has served as President of the Population Association of America and the Association of Population Centers, and has been active on national advisory and study committees of the Census Bureau (CSAC), the National Science Foundation (SBE and ACCI), and the National Academy of Sciences (BRDI).

Catherine A. Fitch is Associate Director of the Institute for Social Research and Data Innovation (ISRDI), the home of the Minnesota Population Center and IPUMS data projects. She is also Co-Director of the Minnesota Research Data Center (MnRDC). She received her PhD in history (2005) and a Master of Public Policy (2001) from the University of Minnesota. At MPC and ISRDI, Fitch has been involved in the creation of several of the largest social-science databases, including IPUMS (USA and International), National Historical Geographic Information System (NHGIS), Terra Populus, and the IPUMS-Ancestry database of complete-count American census data from 1790-1940. She has used her experience with social science data infrastructure to fund and build the MnRDC, a Census Bureau Research Data Center (RDC) providing access to restricted and confidential data. Her own research focuses on historical demography and marriage formation in the United States from the nineteenth century to the present. Fitch served on the ICPSR Council for 4 years from 2009 to 2013.