Big Microdata from the U.S. Census

Steven Ruggles
Catherine Fitch
Matthew Sobek
University of Minnesota

This paper describes projects to create and disseminate integrated microdata with approximately 1.8 billion person-records derived from census enumerations of the U.S. from 1790 to 2012. The availability of massive collections of information on U.S. population characteristics over two centuries 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.

1790-1930 Census Microdata

Over the past three decades, Ancestry.com (the largest for-profit genealogical company) and FamilySearch (a non-profit genealogical organization operated by the LDS Church) each transcribed the records from every surviving public U.S. census. Between 1980 and 2011, FamilySearch used crowdsourcing technology to digitize all the available U.S. Census data. The project attracted over 100,000 volunteers; to maximize accuracy, two volunteers independently keyed each entry, and a third volunteer arbitrated discrepancies. Ancestry.com independently began digitizing U.S. censuses in August 2000, using offshore data-entry vendors to minimize cost. By June 2006, Ancestry had digitized the names and basic demographic characteristics of all persons in the U.S. censuses then publicly available, also spanning the period from 1790 to 1930 (Ancestry.com 2006).

In July 2008, FamilySearch and Ancestry reached an agreement to merge their indexes for the 1790-1930 censuses to improve accuracy (Ancestry.com 2008). Under the agreement, FamilySearch combined their crowdsourced data with the Ancestry datasets and reconciled all discrepancies manually using third-party arbitrators. Thus, the 1790-1930 census records have now been entered three times—twice by FamilySearch volunteers and once by Ancestry's commercial vendors. The two organizations completed the reconciliation project in 2012, and both now offer the enhanced data through their web-based look-up systems.

In March 2013, the University of Minnesota signed an agreement with Ancestry.com that will make the merged data collections available for scientific research and educational purposes, in exchange for data, metadata, and technical infrastructure developed by the IPUMS and NHGIS projects. With some 680 million person-records, the 1790-1930 census collection represents one of the largest-scale data-entry efforts ever undertaken. Ancestry.com and FamilySearch devoted approximately 22 million hours to the transcription of information describing 618 million persons, the equivalent of over 10,000 person-years of effort. The data-entry cost to replicate the collection in the U.S. would be about $420 million.[1]

[1] This estimate covers the costs of dual keying only; data cleaning, checking, and reconciling two copies would incur additional expense. The cost estimate assumes the average Ancestry.com keying rate and the U.S. average salary for data-entry keyers according to the

The 1790-1930 microdata collection does not include every variable enumerated. The digital files for all census years include a core set of variables valuable for demographic research, including geographic location, age, sex, and race, as well as name. Birthplace information is available in all but a few of the early years, and from 1880 forward the data include marital status, the relationship of each individual to the household head, and the birthplace of each individual's mother and father, allowing the identification of second-generation Americans. Other key variables—such as year of immigration, duration of marriage, literacy, occupation, children ever born, children surviving, and disability—are available sporadically. Because of a separate collaboration with FamilySearch, however, we will soon have all variables from the 1850 census, including the slave and mortality schedules.

1940 Census Microdata

In 2012, the University of Minnesota launched a collaboration with Ancestry.com to digitize the entire census of 1940, which had just been released to the public by the National Archives and Records Administration. Ancestry.com had planned to digitize the basic census questions needed by genealogists: name, age, sex, marital status, and birthplace. The University made an agreement with Ancestry.com to split the additional costs needed to digitize the entire census form, and make the full census freely available for scientific research and education. With 136 million person records and 70 variables, the 1940 census database is the largest data collection from a single census ever made freely accessible for scientific research.

The database provides the earliest information available on educational attainment, migration status, labor force status, wage and salary income, hours worked per week, and weeks worked last year. Accordingly, it will provide the baseline for critical analyses of social and economic change. Researchers can link the 1940 census to recent health surveys, administrative records, and the national death index to the 1940 database, allowing study of the impact of early life conditions—including socioeconomic status, parental education, and family structure—on later health and mortality. The data cover the entire population with full geographic detail, providing contextual information on childhood neighborhood characteristics, labor-market conditions, and environmental conditions.

Restricted Microdata, 1960-2012

The Census Research Data Centers (RDCs) operated by the Census Bureau's Center for Economic Studies house complete decennial census microdata of the U.S. for 1970 through 2000, including both short-form and long-form records. Thanks to a nearly-completed Minnesota Population Center project to restore the 1960 census, complete long-form data covering 25% of the 1960 population will soon be available through the RDCs (Ruggles et al. 2011). The RDCs also house complete American Community Survey data covering 47 million persons, and are adding about 5.4 million persons per year. In all, the microdata housed in the RDCs currently describe 1.05 billion persons with full geographic identification down to the block level (Ruggles 2013).

The Minnesota Population Center is currently converting these microdata into IPUMS format, which will greatly simplify cross-temporal analysis. We have nearly completed the metadata and software needed to convert the decennial censuses for 1960 through 2000, and expect to complete testing and validation by the end of 2014. We will then begin work on the ACS data.

Bureau of Labor Statistics (2011). I estimate 57.6 billion keystrokes for dual keying, including U.S. outlying areas such as Puerto Rico as well as the slave and mortality schedules.

Data Processing and Dissemination

Converting the 1790-1940 census transcriptions into a format suitable for scientific analysis will require innovative approaches to big data processing. The scale of the material precludes the use of the manual editing and classification procedures we used for previous census projects; instead, we are developing and implementing fully automated data processing tools. We must carry out the following tasks: (1) classify and code geographic locations to be compatible with categories used in the published census returns; (2) assess completeness and accuracy of the data transcription; (3) convert alphabetic string data into numeric categories that are comparable over time; (4) employ new data cleaning software to identify and correct common enumeration and transcription errors; (5) develop documentation, including full descriptions of data processing methods, detailed analysis of comparability issues, and comprehensive machine-processable metadata; (6) incorporate the data into the IPUMS data access system for free dissemination to the scientific community; and (7) implement secure data protection and preservation policies. Our team of highly experienced researchers has exceptional proficiency in large-scale data creation, integration, and dissemination, and will leverage cutting-edge technology to process an unprecedented volume of data at reasonable cost. We have in hand all the funding that will be needed for these activities, but we are just beginning work. As will be detailed in the full paper, we will release subsets of the data annually from 2014 to 2018.

All numerically coded 1790-1940 data will be openly accessible to the public without restrictions. For users who do not need name information, access will be just like the existing U.S. IPUMS data—anyone will be able to register, create an extract, and immediately download data. For those who need access to the names, we must actively manage data access. The digitized names are the most valuable assets of both Ancestry and FamilySearch, and we have an obligation to our donors to ensure that those names are secure, to protect legitimate commercial interests. There is no privacy concern, because the source data are in the public domain, but Ancestry and FamilySearch own the *digital transcription* of the information. Accordingly, we have agreed to screen requests for access to names, and verify that there is a legitimate scientific need for the information. Because Ancestry wants to limit the number of copies of the entire database in circulation, the license agreement specifies that we may disseminate up to 10 copies each year of the entire database with all names. If we receive more than 10 requests for the entire database in a given year, our steering committee will prioritize the requests on the basis of scientific merit. Our agreement also allows us to disseminate *unlimited* copies with names of subsets of the data that include up to 20% of the population. We will require researchers who want to access the names to sign an agreement guaranteeing data security.

Access to the recent data (1960-2012) is considerably more difficult. Prospective users must prepare a detailed proposal to the Census Bureau documenting both scientific merit of the research and the benefits to the Census Bureau. Once approved, they must undergo security clearance and confidentiality training to obtain Special Sworn Status. Finally, they must conduct the research in one of 18 Research Data Centers located around the country, most of which have significant charges for access. In the short run, analysis will be further hindered by inadequate computing capacity and awkward file formats, but we expect that these issues can be resolved. We eventually hope to bring a version of the IPUMS data access system into the RDC environment, which would greatly simplify access to these massive files. In the long run, it should become possible to import the complete historical enumerations of 1790-1940 into the RDC, thereby supporting research requiring billions of records of fine-grained spatiotemporal data spanning more than two centuries.

Funding Sources

Development of Big Data from the U.S. Census is supported by the following grants:
NICHD R01HD078322 “Big Data for Population Research”
NSF SES1155572 “Infrastructure for Population Analysis”
NICHD R01HD073967 “Baseline Socioeconomic Microdata for Population and Health Research”
NIA R01AG041831 “Microdata for Analysis of Early Life Conditions, Health, and Population”
NICHD R01 HD43392 “IPUMS Redesign”
NICHD R01HD060676 “Baseline Microdata for Analysis of U.S. Demographic Change”

References

Ancestry. 2006. Press release: Ancestry.com digitizes entire u.s. federal census collection from 1790-1930. http://corporate.ancestry.com/press/press-releases/2006/06/ancestry.com-digitizes-entire-u.s.-federal-census-collection-from-1790-1930/

Ancestry.com. 2008. Press release: FamilySearch and Ancestry.com team to publish new images and enhanced indexes to the U.S. Censuses. http://corporate.ancestry.com/press/press-releases/2008/07/familysearch-and-ancestry.com-team-to-publish-new-images-and-enhanced-indexes-to-the-u.s.-censuses/

Bureau of Labor Statistics. 2011. *Occupational employment and wages, May 2011*. 43-9021 Data Entry Keyers. http://www.bls.gov/oes/current/oes439021.htm.

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

Ruggles, S. 2013. Big Microdata for Population Research. *Demography* DOI 10.1007/s13524-013-0240-2 (Published online 7 September 2013).