ADEP WORKING PAPER SERIES

The Decennial Census Digitization and Linkage Project

Katie R. Genadek
U.S. Census Bureau

J. Trent Alexander
University of Michigan

Working Paper 2019-01

September 2019

Associate Directorate for Economic Programs
U.S. Census Bureau
Washington DC 20233

The Decennial Census Digitization and Linkage Project

Katie R. Genadek, U.S. Census Bureau, katie.r.genadek@census.gov
J. Trent Alexander, University of Michigan, jtalex@umich.edu
ADEP Working Paper 2019-01
August 2019

Abstract
The Decennial Census Digitization and Linkage project (DCDL) is an initiative to produce linked restricted microdata files from the decennial censuses of 1960 through 1990. This paper provides background on the decennial census files and the previous work leading to the DCDL project. The proposed work plan is described in detail, as well as the dissemination strategy for the resulting linked data. When combined with existing linkages between the censuses of 1940, 2000, 2010, the soon-to-be public 1950 census, and the future 2020 census, the DCDL project will provide the final component in a longitudinal data infrastructure that covers most of the U.S. population since 1940.

Keywords: census data; census history; data capture

I. Introduction

The Decennial Census Digitization and Linkage project (DCDL) is an initiative to produce linked restricted microdata files from the decennial censuses of 1960 through 1990. When combined with existing linkages between the censuses of 1940, 2000, 2010, the soon-to-be public 1950 census, and the future 2020 census, DCDL will provide the final component in a longitudinal data infrastructure that covers most of the U.S. population since 1940. As a multi-purpose statistical tool, the DCDL will further the Census Bureau's mission to provide high quality data on the U.S. population, and it will be used by the Census Bureau to improve the data collected by Census Bureau surveys. The resulting data resource will expand our understanding of population dynamics in the U.S. far beyond what is currently possible, providing transformational opportunities for research, education, and evidence-building across the social, behavioral, and economic sciences.

DCDL will build upon the Census Bureau's extensive longitudinal data infrastructure, which already includes major national surveys going back to 1973, administrative records going back to the 1990s, and the decennial censuses from 1940, 2000, and 2010. These restricted data are used by Census Bureau employees and are available to approved researchers around the U.S. via the Federal Statistical Research Data Centers (FSRDCs). The linked census data are accessible through clearly-defined access protocols and are currently being used by more than fifty researchers across the thirty FSRDC locations.

The Census Bureau already maintains and makes available nearly-complete microdata files for 1960 through 1990. Those files include all variables *other than* respondent names, which were never digitized due to the cost of transcription. Without names, however, we cannot link individuals in these censuses over time. DCDL will manage the difficult and resource-intensive task of capturing the names from the census images stored on microfilm and adding respondent names to the 1960-1990 census microdata files in order to incorporate new linked data files into the Census Bureau holdings available to FSRDC researchers.

The power of these linked data will be not only in their scope (they will cover the period from 1940-2020) and scale (they will include most of the U.S. population), but also in the ease with which additional files can be linked to these data for analysis. For instance, nearly all survey data and administrative data held at the Census Bureau are routinely integrated into the agency's data linkage infrastructure shortly after they are created or acquired. In addition, approved

researchers will be able to bring their own data into the FSRDCs to have them linked and available for use with the longitudinal census data.

The DCDL project has three main objectives:

1. *Digitize and recover respondent names from microfilmed decennial census manuscripts of 1960, 1970, 1980, and 1990*. We will use high-speed microfilm scanners to create more than 500 million digital images from the 258,000 microfilm reels, and then we will use Optical Character Recognition (OCR) processes to create machine-readable respondent names from these images.
2. *Attach recovered names to existing microdata files.* The Census Bureau has microdata files from the 1960 through 1990 censuses that include all variables other than respondent names. In order to append names to the correct record in the existing microdata, we will capture and make use of several additional fields from the census form images from each year.
3. *Link the 1960-1990 censuses together and to the 1940, 1950, 2000, 2010, and 2020 censuses.* These linkages will utilize well-established methods that have been used to build an infrastructure that already contains censuses, surveys, and administrative data before and after this period. Individual names will be replaced with unique identifiers that permit researchers to trace individuals over time. The resulting restricted-use linked files will form the core of a statistical infrastructure documenting most of the U.S. population since 1940.

II. Background

The U.S. Census Bureau has been innovating in the field of automated record linkage for the past five decades (Jaro 1972; Jaro 1989; Winkler 1994, 1999, 2000). Current and former Census Bureau employees, including members of the DCDL team, developed unprecedented capabilities for large-scale linkage of restricted data (Massey and O'Hara 2014; Wagner & Layne 2014; Johnson et al. 2015; Alexander et al. 2017, Massey et al. 2018). The Census Bureau's production record linkage systems operate by matching any individual-level data to a large composite of administrative records. Valid links are assigned a unique Protected Identification Key (PIK), and Personally Identifiable Information (PII) is removed from the records. The PIKs facilitate linkage to any other file that has been assigned PIKs. The census, survey, and administrative data in this infrastructure are primarily from Census 2000 to the present. Most of these twenty-first century files contain PII

that is sufficient to link between 85-99% of cases using probabilistic linkage techniques (Wagner & Layne 2014) with minimal error (Layne et al. 2014).

Older census files pose greater linkage challenges than modern data files, largely because of the limited PII (Massey 2017; Massey et al. 2018, Ruggles et al. 2018). In addition to linking modern data, we have performed extensive research on the recovery and linkage of historical data, including the decennial censuses in the current project (Ruggles et al. 2011; Alexander et al. 2014; Massey & O'Hara 2014). Alexander and Genadek managed the 1960 Census Data Restoration Project, a collaboration between the Census Bureau and the University of Minnesota to recover from microfilm geographic areas that were missing from the "long form" of the 1960 census microdata (Ruggles et al. 2011). Alexander's team at the Census Bureau subsequently carried out a successful pilot study to assign PIKs to a small set of names recovered from the 1960 census (Massey 2014). In 2015, the Census Bureau also added the complete 1940 census data to the linked data infrastructure. Working with researchers from several universities, Alexander and Genadek modified the Census Bureau's linkage systems to accommodate the variables available in the 1940 census. In the beta version of the linked 1940 census data, PIKs were assigned to 70% of children and 40% of the total population in 1940 (Alexander et al. 2014, 2017; Massey et al. 2018) and the Census Bureau is working to increase these linkage rates.

These infrastructure-building projects demonstrate that twentieth-century census data can be recovered from microfilm, that the Census Bureau's linkage systems can be adapted for older data, and that the DCDL team is uniquely qualified to manage this challenging task. Given that the currently available 1960-1990 census microdata do not contain names, these files have not received PIKs and are not included in any linked data resource. The currently available longitudinal data include the 1940 census, Census 2000, the 2010 census, and survey data samples dating from the mid-1970s. While this longitudinal infrastructure is generating path-breaking scholarship, the addition of linked data between 1940 and 2000 will enable new possibilities for innovative research and data products (Ferrie et al. 2016, Liebler et al. 2016, Liebler et al. 2017, Alexander et al. 2017).

Social scientists have been calling for the use of linked census, survey and administrative data since at least the 1970s (Ruggles & Ruggles 1974), though it has been widely acknowledged that the cost of this infrastructure-building was well outside the bounds of ordinary funding mechanisms (Okner 1972). More recently, academic researchers and government practitioners

have revisited and reinvigorated this idea, conducting a series of workshops, standing groups, and pilots to prepare to build a large longitudinal panel of social and economic data for the United States.

In 2012, with support from the National Science Foundation (NSF) and the National Academy of Sciences, the National Research Council invited a group of social mobility experts to prepare articles proposing new data sources for studying social mobility. Presented at a June 2013 workshop at the National Academies and subsequently published, several of the commissioned papers came to the same conclusion: the best way to create a new dataset for studying social mobility was to link already-collected census and survey data over the twentieth century (Grusky et al. 2015; Johnson et al. 2015; Warren 2015).

A core team from the NSF-funded workshop formed the basis of a new National Academies Standing Committee called "Creation of the American Opportunity Study (AOS)." AOS team set out to define the research needs and technical requirements for building an infrastructure of census data, survey data, and administrative data to facilitate studies of social and economic mobility. In May 2016, the AOS Standing Committee invited researchers from a broad range of disciplines to discuss challenges in creating an infrastructure and to build evidence that such a tool was needed across the social sciences. The workshop included presentations from researchers in sociology, economics, anthropology, statistics, public policy, criminology, survey practice, and the federal government. These presentations are summarized in the conference proceedings (National Academies 2016).

The AOS Standing Committee helped to document the research community's broad support for large-scale longitudinal tools, ultimately making clear that the demand for this type of resource goes far beyond research in social and economic mobility. In the past few years, for instance, scholars have argued that such a data system is needed to investigate education (Bloome et al. 2018), family development (O'Hara et al. 2016), political processes (Brady et al. 2015), multi-generational change (Song and Campbell 2017), and government program evaluation and evidence-building (Abraham et al. 2018).

Concurrent with the National Academies organizational activities, AOS leaders collaborated with a Census Bureau team led by Alexander to define techniques and costs for recovering names from the 1990 census. With funding from the Carnegie Foundation through the National Academies, the 1990 Name Recovery Pilot (NRP) created digital images from a sample

of 1990 census microfilm, hand-keyed and verified "truth data" from those images, supported two vendors' attempts to digitize the names and additional variables by conducting OCR on the images, and linked the recovered names and data into the Census Bureau infrastructure. The best OCR vendor's results provided machine-readable names that exactly matched the hand-keyed data 82% of the time for last names and 75% of the time for first names. The 1990 NRP team carried out record linkage with a set of cases for which they had both OCR-provided names and hand-keyed names. When linked to an administrative records composite and assigned PIKs, OCR-provided names and hand-keyed names produced the same PIK 87% of the time. The accuracy of linkages made with OCR-based names can be further improved by post-linkage verification of individual characteristics against other files with PIKs, and by a tightening of linkage parameters. In addition to proving that the data capture and linkage of these records is possible, the NRP provided invaluable experience for defining the methods and workflows needed for DCDL.

Since the completion of the 1990 NRP, we have continued to work closely with the AOS team and the Census Bureau operations staff to prepare for full-scale production of name recovery on the 1960 through 1990 censuses. With funding from the University of Michigan, Stanford University, and the Hewlett Foundation, we have inventoried and viewed samples of all microfilm reels, prepared preliminary code to match recovered names to existing microdata files, and gained access to contemporaneous administrative records to facilitate the linkages necessary for DCDL.

III. Proposed Work Plan

The key goal of this project is to create linked data from the 1960-1990 censuses. The 1950 census will become public in 2022 and the 2020 census data will also be available around that time, resulting in linked data from 1940-2020. To this end, DCDL has three main objectives: (1) recover respondent names from the 1960-1990 decennial censuses using high-speed microfilm scanning and handwriting recognition technology, (2) attach respondent names to otherwise complete microdata files held at the Census Bureau, and (3) assign unique identifiers to the data so that individuals can be linked from file to file following the removal of the PII.

Objective 1. Recover respondent names from the 1960 through 1990 decennial censuses.

In order to link these censuses over time, we need to recover respondent names. The conventional method for digitizing names from census and survey data is enter them manually. The Census Bureau manually entered names from Census 2000 and the 2010 Census, and genealogy

companies and academic institutions continue to use manual methods to digitize files from older censuses around the world. With over 850 million respondents between 1960 and 1990, the cost of traditional data entry is not feasible.

The only cost-effective means of digitizing information in these censuses is by using automated Optical Character Recognition (OCR) to capture names and additional variables to facilitate further linkage. The OCR-based name recovery involves two distinct processes: digitizing original census images and conducting OCR on the digitized files. Even using well-tested methods and automating tasks whenever possible, the digitization of census images is the most labor-intensive part of the DCDL project.

Digitizing original census images. The original census manuscripts are stored as more than half a billion photographic images on 258,000 microfilm reels. The reels are housed at the Census Bureau's National Processing Center in Jeffersonville, IN. Because all of these files are still protected by the confidentiality provisions of Title 13 of the U.S. Code (and will remain so for decades), all digitization and linkage work must take place in secure Census Bureau facilities and be carried out by Census Bureau staff and contractor. These security provisions do contribute to project cost, though DCDL's scanning team will draw from the same highly experienced data processing staff who are the intake point for all census and survey responses for the agency.

Our methods for capturing the names from the 1960-1990 censuses build upon well-established processes that we honed in the course of the 1960 Census Data Restoration Project and the 1990 NRP. Each 100-foot microfilm reel can be scanned and digitized in just over 20 minutes, including time for necessary record-keeping and quality control. A technician can operate two scanners simultaneously, and therefore can scan about 5 reels per hour, or 9,000 reels per year. With eleven technicians working full-time on the project, we estimate that all reels will be scanned within 30 months.

Through our pilot projects, we worked with two scanning software specialists to establish optimal scanner and image settings to accommodate OCR. Each reel contains around 1,500-2,000 images. With optimized settings, these images comprise a total of two gigabytes of data per reel. When complete, the digitized scans of all images from the 1960-1990 censuses will fill 500 terabytes of secure storage.

Conducting OCR on digitized census images. After we have created digital images of the census records, we will use OCR to capture names and several other variables. We will manually enter a

set of "truth data" from each census. A portion of that data be used to train OCR algorithms, and a portion will be held back and used to measure the accuracy of the OCR output. In the 1990 NRP project, we held a competition between two vendors, one of which was the University of Southern California's Information Sciences Institute (USC ISI), a leader in multi-lingual OCR and document image processing. The ISI team produced the highest quality OCR results.

While the USC ISI team was extraordinarily successful in our 1990 census pilot project, we need to allow for the possibility that OCR technology may have advanced significantly since that time. For instance, genealogical companies are testing novel techniques to produce high-quality output on images from public documents. We will investigate new OCR technology and vendors, and develop a contract with the best option at the time. All contractors will have Census Bureau Special Sworn Status, which requires a moderate background check and swearing to protect of confidentiality of the data for life, and perform the work in a secure environment.

Objective 2. Attach recovered names to existing microdata files.

After we have recovered names from the original census images, we will need to attach those names to the existing microdata files from 1960 through 1990 in order to assign unique and confidential identifiers to the data. While these linkages are not highly complex, they rely on our ability to capture additional variables from the census images. In 1990, for instance, each census form has a unique 11-digit code that allows us to attach recovered names to existing microdata files. Since there will be a small amount of error in the OCR's ability to capture this 11-digit code, we will also collect each household head's age and sex data from the images. Capturing those additional variables will allow us to validate and measure the quality of each linkage from the census images to the census microdata.

This process will be more complex for 1960, 1970, and 1980, because there is not a single identification variable for each image in those years. The microfilmed records are organized into Enumeration Districts (EDs). EDs are census-defined geographic areas that typically contain several hundred people. On each reel, the beginning of a new ED is identified with a "Breaker Sheet" image containing the ED number. ED numbers are indicated via "bubbles" that can be read by Optical Mark Recognition (OMR) software, similar to that used in the 1960 Data Restoration Project. Following the ED Breaker Sheet, each census manuscript image contains a bubbled page number value. The page number, along with the ED number from the Breaker Sheet, is used to associate each image with a microdata record. Our previous experience has shown that a small

number of these page number values are not unique within Enumeration Districts in 1960 and 1970. For those years, we will use household size and age to make the final determination where questionable matches arise. The OMR processes will capture the additional information that we need in order to make and validate the linkage of the name to the microdata file.

Objective 3. Create linkages over time.

DCDL will adapt well-established techniques to conduct record linkage across the 1960-1990 censuses, using a modified version of the Census Bureau's production record linkage processes (Alexander et al 2017; Massey et al. 2018, Massey 2017, Wagner and Layne 2014). Similar to completed linkages for the 1940, 2000, and 2010 censuses, our approach will be to link each census respondent to their record in a composite of administrative data.

The key components in administrative records composite are (1) demographic information from Social Security Card application data from the Social Security Administration's "Numident" database and (2) year-specific place-of-residence information from annual income tax returns. We plan to use individual income tax returns from tax years 1969, 1979, and 1989; those returns were filed in the spring of 1970, 1980, and 1990, at nearly the same time as the censuses were conducted. Tax returns from 1960 were not retained, so that linkage will rely entirely on information from Social Security Card applications.

Table 1 shows the linkage keys that will be used for each census year. In all years, linkage will rely on name, age, sex, state or country of birth, and—for children living with their parent(s)—parents' names. For 1970-1990, linkage will additionally rely on place-of-residence information from tax returns. Whenever a link is made from a census to the administrative records composite, the record is assigned a unique PIK. Respondent names are removed from the data file following PIK assignment. Any files with PIKs can then be easily linked to other files with PIKs in the census infrastructure, without the use of names or other identifying information.

Table 1. Available matching characteristics, 1960-1990

	Census Microdata Records		Administrative Records Composite	
Year	Geography	Characteristics	Source	Characteristics
1990	Street Address	Name, Age, Sex, Marital Status, State or Country of Birth*, Parents' Names**	Social Security Card Applications and 1989 Tax Returns	Name, Birthdate, Sex, Marital Status, State or Country of Birth, Parents' Names, Address
1980	Census tract (metro) or county subdivision (rural)	Name, Age, Quarter of Birth*, Sex, Marital Status, State or Country of Birth*, Parents' Names**	Social Security Card Applications and 1979 Tax Returns	Name, Birthdate, Sex, Marital Status, State or Country of Birth, Parents' Names, Address
1970	Census tract (metro) or county subdivision (rural)	Name, Age, Quarter of Birth, Sex, Marital Status, State or Country of Birth*, Parents' Names**	Social Security Card Applications and 1969 Tax Returns	Name, Birthdate, Sex, Marital Status, State or Country of Birth, Parents' Names, Address
1960	Not used in linkage	Name, Age, Year of Birth, Quarter of Birth, Sex, Marital Status, State or Country of Birth*, Parents' Names**	Social Security Card Applications	Name, Birthdate, Sex, Marital Status, State or Country of Birth, Parents' Names

* Available only in the long-form
** Available only for respondents living with parents

In addition to the basic linkage strategy that we will use for all cases, we will also use variables that are only available in the census "long form," a more detailed questionnaire that is administered to approximately 16%-25% of the population each year. While these variables are not necessary for most linkages, they may help with assigning PIKs to cases that we are not able to link initially.

The application of PIKs to the 1940 census, which relied on a subset of DCDL's linkage keys, produced PIKs for 40% of adults and 70% of children (Massey et al. 2018). Each of the 1960-1990 censuses has additional features that will yield even higher linkage rates than those obtained for 1940. As shown in Table 1, for the 1970-1990 linkages, we will make use of residential geography. In 1990, this includes full street address. In 1970 and 1980, when the census data does not have full street address, we will map street addresses from 1969 and 1979 tax records to geographic areas identified on each year's census data, relying on boundary files from the National Historical Geographic Information Systems (NHGIS). When assigning PIKs to the 1960-1980 censuses, we will also use the quarter of birth variable. Used in conjunction with age, this variable allows for a finer-grained match to the date of birth variable in the administrative records

composite file. Census Bureau tests from a pilot project to link the 1960 census showed that this variable significantly improved the assignment of PIKs (Massey 2017).

Even with high-quality data having all necessary matching variables, PIKs are subject to a small amount of linkage error, and there are issues with coverage and bias (Bond et al. 2014, Layne et al. 2014). In recent census data, more than 90% of the population receive PIKs; research suggests that nearly all of these assigned PIKs are accurate (Layne et al. 2014). Still, PIKs have been shown to under-represent the highly mobile population, those with poor spoken English, non-citizens, and other groups. Linkage rates are lowest in the Southwest, and highest in the upper Midwest. Because the pre-2000 administrative records composites are built from fewer sources than more recent composites (and only from one source in 1960), PIKs we generate for 1960-1990 could be subject to additional bias. To assess these issues, we will replicate each year's linkage approach with Census 2000 data, and then compare PIKs generated with the full information available in the Census 2000 data to PIKs generated with the more limited information that was available in the 1960-1990 censuses. These tests will allow us to make assessments of the additional bias and error inherent in the methods we use for the older censuses.

Project Timeline.

We estimate the DCDL project will take six years to complete. The project will begin with the microfilm scanning and the manual entry of "truth data", which will take approximately three years. The development and application of OCR processes will take place concurrently with scanning and will continue through the fourth year. The linkage of names to currently held microdata files will begin during the second year and will be completed shortly after the OCR is finished, at the end of year 4. The development of longitudinal record linkage strategies will take place iteratively over years 2 through 4 and will be carried out on the 1960-1990 censuses during year 5. The final year of the project will focus on the quality assessment and the creation of documentation and dissemination tools.

IV. DCDL Output and Dissemination

At the conclusion of the project, DCDL data will complete a data series including almost the entire population between 1940 and the present. The data will be integrated over time and will be linked to other surveys and administrative records held at the Census Bureau. Using de-identified data, scholars and Census Bureau employees will be able to follow individuals over their lives, from

childhood to old age, and families will be traceable across generations. The DCDL-based data will provide a multi-purpose data infrastructure that will be maintained by the Census Bureau to improve the measurement of the U.S. population, and the data will be made accessible to researchers through the FSRDC network.

The methods used to create the DCDL data are flexible enough to incorporate additional data sources, thereby continuously enabling new types of data-intensive research. Long-term linked census data will multiply the value of the most important surveys and infrastructure projects in the social sciences, such as the Panel Study of Income Dynamics, the Wisconsin Longitudinal Study, the Current Population Survey, the American Community Survey, and the Health and Retirement Study. Furthermore, individual researchers with identifiable data on program participants, experiments and interventions, or from smaller surveys will be able to bring those data into the secure FSRDC environment, where Census Bureau staff can merge the data with longitudinal census files. In all of these ways, the DCDL data will enable transformational research by providing demographic and socio-economic dynamics on individuals and their families over decades and generations.

DCDL microdata are protected by the confidentiality provisions of Title 13 of the U.S. Code and will be retained in the enterprise research computing environment managed by the Census Bureau and made available through the FSRDCs. As is the case for all FSRDC resources, DCDL data will not contain names or SSNs. Researchers will use anonymized PIKs to link de-identified records over time. As the largest agency in the Federal Statistical System, the Census Bureau has a long track record of retaining and providing access to its own high-value microdata along with data from other agencies. The Census Bureau has supported the FSRDCs since forming the network in 1992 to provide researchers with access to non-public microdata. The FSRDCs currently support hundreds of researchers using restricted data produced by the Census Bureau and other federal agencies. The FSRDCs have well-established protocols for researchers outside of the Census Bureau to obtain access to restricted data. In 2018, following a successful pilot program led by DCDL team leads, the Census Bureau began allowing outside researchers to propose and conduct projects using linked decennial census data. The use of the linked census data requires approval only from the Census Bureau, making access more straightforward than it is for some other federal agency data held in the FSRDCs. In order to ensure the broadest dissemination of these data resources, all project metadata and documentation will be publicly available at the

Census Bureau Data Repository at the University of Michigan's Inter-university Consortium for Political and Social Research (ICPSR).

V. Conclusion

For the past five decades there has been considerable community interest in creating a large longitudinal data resource that integrates census data, survey data, and administrative records. Due to the development of new data resources and data processing techniques, this vision has moved closer to reality in the past few years. We have conducted numerous pilot projects to develop the required methods and operational experience to complete the digitization of names from the 1960-1990 censuses, attach unique identifiers to the respondents in this data, and link the census data over time. These efforts have been supported by robust partnerships between universities, foundations, federal agencies, and the FSRDC network.

The DCDL project will allow researchers to use millions of linked cases over decades, permitting the longitudinal study of the U.S. population at a finer level than has ever been possible. Research conducted with the data resulting from DCDL will transform our understanding of intergenerational, social, and geographic mobility, as well as life-course transitions and trajectories. The resulting data series will serve both academic social sciences and the evaluation of government programs and policies.

DCDL will complete a large gap in the Census Bureau's data linkage infrastructure, fulfilling the Census Bureau's mission to provide high quality data about the population. The Census Bureau, as well as researchers around the U.S., will use this resource intensively. The files created through the current project will make the 1940-2020 linked decennial census data a core element of the nation's statistical infrastructure.

VI. References

Abraham, K. G., Haskins, R., Glied, S., Groves, R. M., Hahn, R., Hoynes, H., & Wallin, K. R. (2018). The Promise of evidence-based policymaking: Report of the commission on evidence-based policymaking.

Alexander, J. T., Gardner, T., Massey, C. G., & O'Hara, A. (2014). Creating a longitudinal data infrastructure at the Census Bureau. *Work. Pap., Center for Administrative Records Research and Applications, US Census Bureau*.

Alexander, J. T., Leibbrand, C., Massey, C., & Tolnay, S. (2017). Second-Generation Outcomes of the Great Migration. *Demography, 54*(6), 2249-2271.

Bond, B., Brown, J. D., Luque, A., & O'Hara, A. (2014). The nature of the bias when studying only linkable person records: Evidence from the American Community Survey. *Center for Administrative Records Research and Applications Working Paper, 8*.

Brady, H. E., Schlozman, K. L., & Verba, S. (2015). Political mobility and political reproduction from generation to generation. *The ANNALS of the American Academy of Political and Social Science*, 657(1), 149-173.

Ferrie, J., Massey, C. & Rothbaum, J. (2016). Do Grandparents and Great-Grandparents Matter? Multigenerational Mobility in the US, 1910-2013. *National Bureau of Economic Research Working paper No. 22635*.

Grusky, D. B., Smeeding, T. M., & Snipp, C. M. (2015). A new infrastructure for monitoring social mobility in the United States. *The ANNALS of the American Academy of Political and Social Science, 657*(1), 63-82.

Jaro, M. A. (1972, May). UNIMATCH: a computer system for generalized record linkage under conditions of uncertainty. In *Proceedings of the May 16-18, 1972, spring joint computer conference* (pp. 523-530). ACM.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association, 84*(406), 414-420.

Johnson, D. S., Massey, C., & O'Hara, A. (2015). The opportunities and challenges of using administrative data linkages to evaluate mobility. *The ANNALS of the American Academy of Political and Social Science, 657*(1), 247-264.

Layne, M., Wagner, D., & Rothhaas, C. (2014). Estimating record linkage false match rate for the Person Identification Validation System. *Center for Administrative Records Research and Applications Working Paper, 2*.

Liebler, C. A., Porter, S. R., Fernandez, L. E., Noon, J. M., & Ennis, S. E. (2017) America's Churning Races: Race and Ethnic Response Changes between Census 2000 and the 2010 Census. *Demography* 54(1): 259-284.

Liebler, C. A., Bhaskar, R., & Porter, S. R. (2016). Joining, Leaving, and Staying in the American Indian/Alaska Native Race Category Between 2000 and 2010. *Demography* 53(2): 507-540.

Massey, C. G. (2014). *Creating linked historical data: An assessment of the Census Bureau's ability to assign protected identification keys to the 1960 Census* (No. 2014-12). Center for Economic Studies, US Census Bureau.

Massey, C. G., & O'Hara, A. (2014). *Person Matching in Historical Files using the Census Bureau's Person Validation System* (No. 2014-11). Center for Economic Studies, US Census Bureau.

Massey, C. G. (2017). Playing with matches: An assessment of accuracy in linked historical data. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *50*(3), 129-143.

Massey, C. G., Genadek, K. R., Alexander, J. T., Gardner, T. K., & O'Hara, A. (2018). Linking the 1940 US Census with modern data. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *51*(4), 246-257.

National Academies of Sciences, Engineering, and Medicine (2016). *Using Linked Census, Survey, and Administrative Data to Assess Longer-Term Effects of Policy: Proceedings of a Workshop*. Washington, DC: National Academies Press.

O'Hara, A., Shattuck, R. M., & Goerge, R. M. (2017). Linking Federal Surveys with Administrative Data to Improve Research on Families. *The ANNALS of the American Academy of Political and Social Science*, 669(1), 63-74.

Okner, Benjamin A. (1972). Conducting a New Data Base from Existing Microdata Sets: The 1966 MERGE File. *The Annals of Economic and Social Measurement*, 1/3: 325-337.

Pfeffer, F. T., & Hertel, F. R. (2015). How has educational expansion shaped social mobility trends in the United States?. *Social Forces*, 94(1), 143-180.

Reeves, R. V. (2016). How Will We Know? The Case for Opportunity Indicators. In *The Dynamics of Opportunity in America* (pp. 443-464). Springer, Cham.

Ruggles, N., & Ruggles, R. (1974). A strategy for merging and matching microdata sets. *Annals of Economic and Social Measurement,* 3(2): 353-371. NBER.

Ruggles, S., Schroeder, M., Rivers, N., Alexander, J. T., & Gardner, T. K. (2011). Frozen film and FOSDIC forms: Restoring the 1960 US Census of Population and Housing. *Historical methods*, *44*(2), 69-78.

Ruggles, S., Fitch, C. A., & Roberts, E. (2018). Historical census record linkage. *Annual Review of Sociology*, 44(1): 19-37.

Song, X., & Campbell, C. D. (2017). Genealogical microdata and their significance for social science. *Annual Review of Sociology*, 43, 75-99.

U.S. Census Bureau (2019). Research @ Census. Webpage available at https://www.census.gov/research/. Accessed February 12, 2019.

Wagner, D., & Layne, M. (2014). The person identification validation system: Applying the Center for Administrative Records and Research and Applications' record linkage software. *Center for Administrative Records Research and Applications Report Series (# 2014-01)*.

Warren, J. R. (2015). Potential data sources for a new study of social mobility in the United States. *The ANNALS of the American Academy of Political and Social Science*, 657(1), 208-246.

Winkler, W. E. (1994). Advanced methods for record linkage.

Winkler, W. E. (1999). The state of record linkage and current research problems. In *Statistical Research Division, US Census Bureau*.

Winkler, W. E. (2000). Frequency-based matching in Fellegi-Sunter model of record linkage. *Bureau of the Census Statistical Research Division*, *14*.