Using Administrative Records to Support the Linkage of Census Data: Protocol for Building a Longitudinal Infrastructure of U.S. Census Records

J. Trent Alexander, University of Michigan, jtalex@umich.edu
Katie R. Genadek, U.S. Census Bureau and University of Colorado, katie.r.genadek@census.gov

Acknowledgement of funding sources
This work was supported by the National Science Foundation under grant number 2023639.

Using Administrative Records to Support the Linkage of Census Data: Protocol for Building a Longitudinal Infrastructure of U.S. Census Records

Abstract

This article describes the linkage methods that will be used in the Decennial Census Digitization and Linkage project (DCDL), which is an initiative to build a longitudinal infrastructure of modern United States census data. DCDL is digitizing and creating linkages between nearly a billion records across the 1960 through 1990 U.S. censuses, as well as to already-linked records from the 1940 and 2000 through 2020 censuses. Our main goals in this article are to (1) describe the history of the DCDL and the protocol we will follow to build the linkages between the census files, (2) outline the techniques we will use to evaluate the quality of the links, and (3) show how the assignment and evaluation of these linkages leverages the joint use of routinely collected administrative data and non-routine survey data.

Background

This article describes the linkage methods that will be used in the Decennial Census Digitization and Linkage project (DCDL), which is completing the final four decades of a longitudinal census infrastructure covering the past 170 years of United States history. Conducted every 10 years since 1790, the U.S. Census is a mandatory survey with response rates well over 95% since the mid-twentieth century (King and Magnusson 1995). While the U.S. has no population register, researchers have a long history of using linked census data to study long-term social and economic change. Over the past two decades, historical demographers have built an extraordinary longitudinal infrastructure of publicly-available census data from 1850-1940, including major efforts from the University of Minnesota and the National Bureau of Economic Research (Abramitzky et al. 2020, Helgertz et al. 2020, Ruggles et al. 2021). Similar infrastructures have been built from the publicly-available historical censuses from Canada, England, Norway, Sweden, and Wales (Antonie et al. 2014, Ruggles et al. 2011, Ruggles et al. 2020, Wisselgren et al. 2014). All of these projects have focused on data that is old enough to be made publicly available, and none of these countries currently has an integrated infrastructure of linked records that includes both public historical data and confidential modern data. The DCDL is filling this gap for the U.S., creating linkages between nearly a billion confidential records across the mid-to-late twentieth-century censuses, as well as to already-linked records from the 1940 and 2000 through 2020 censuses. When complete, the digitized and linked data will complete a longitudinal infrastructure that spans the entire period from 1850-2020.

The DCDL began as a series of pilot projects in the 2010s and is currently underway (U.S. Census Bureau 2022). The resulting longitudinal data will become part of a wide range of restricted-use survey data and administrative records that is made available to qualified

researchers working in secure facilities called Federal Statistical Research Data Centers (FSRDCs). Researchers working on approved FSRDC projects can currently use linked data from the censuses of 1940 and 2000 through 2020. Beginning with the 2000 Census, the Census Bureau digitized respondent names in the course of data processing and used those names to facilitate linkage over time. While the agency has microdata files from the 1960 through 1990 censuses, those files have never included respondent names and were therefore never linked over time. We first added historical census data to the linked data infrastructure in 2014, when our team acquired the complete 1940 Census, which was made public in 2012 and digitized by Ancestry.com and the University of Minnesota. We developed methods to link the 1940 Census data to the modern census and survey files and made the linkages available within the FSRDCs. We released the linked 1940 data via a pilot program in 2015 and then to all FSRDC researchers in 2018 (U.S. Census Bureau 2021a).

Even though the currently-available series has a 60-year gap between the 1940 Census and the 2000 Census, the linked census data are already being used intensively by sociologists, economists, demographers, and public health scientists (U.S. Census Bureau 2021b). Ongoing projects are advancing research on topics including Medicaid and mortality (Miller et al. 2021), the impact of reproductive health care on education (Stevenson et al. 2021), the inter-generational determinants and outcomes of twentieth-century mobility (Alexander et al. 2017, Leibbrand et al. 2020, Ferrie at al. 2021), and the demography of mortality in the United States (Finlay and Genadek 2021, Polyakova et al. 2021, Miller et al. 2021). Research conducted with the completed infrastructure will have the potential to further transform our understanding of population aging processes, life-course transitions and trajectories, and the long-term antecedents of health, well-being, and mortality over the last century.

With support from a coalition of funders, DCDL is filling in the gap between the 1940 and 2000 censuses. Our initial work has focused on digitizing respondent names from 1960 through 1990. The censuses from these years collected respondent names, but those names were never digitized and are stored only on microfilm in the respondents' handwriting. Beginning in early 2021, the Census Bureau data processing staff began creating digital images from the 250,000 microfilm reels containing the original census manuscripts. The DCDL's handwriting recognition team has begun using automated methods to extract names, adapting work we completed on a pilot project with the 1990 Census. The DCDL linkage work will begin in early 2023, concurrently with ongoing digitization and handwriting recognition. Upon completion of the DCDL project, all linked data will be available in the FSRDCs.

Our approach to creating linkages between censuses will be to integrate routinely-collected administrative records files with one another, and then use the resulting administrative records composite to facilitate the assignment of linkage keys to the non-routine records from the 1960-1990 censuses. By using administrative records that are proximate to census data in time, we reduce the risk of variables changing over time, and we maximize the use of the particular linkage variables available during each census year. This protocol for the linkage of records on the DCDL project builds on methods used to assign anonymous linkage keys to the most recent decennial census data (Wagner and Layne, 2014), as well as lessons learned from research focused on the linkage of historical census records from 1850-1940 (Massey et al. 2018, Bailey et al. 2020, Helgertz et al. 2022).

Our main goals in this article are to (1) describe the history of the DCDL and protocol we will follow to build the linkages between the census files, (2) outline the techniques we will use to evaluate the quality of the links, and (3) show how the assignment and evaluation of these

linkages leverages the joint use of routinely collected administrative data and non-routine survey data. This work is particularly relevant to researchers conducting or evaluating probabilistic linkages between relatively low-information files, as well to those developing large infrastructure projects creating linkages across centuries of public and confidential population data.

Challenges and Preliminary Studies

The longitudinal linkage of decennial census records involves challenges that will be familiar to anyone who has linked records collected at different times. High-quality personally identifying information (PII) is usually more limited in the past, and the digitization of paper-based records often introduces new potential for increased cost and error. Even when a variable is present and accurate in multiple datasets over decades, linkages need to account for changes in respondents' lives. People may have changed their names, their family composition, their addresses, and even the way they describe where they were born. There is no national identification number in the United States; U.S. Social Security Numbers (SSNs) are a non-universal identifier and have never been requested of census respondents. For this reason, all linkages over time are probabilistic and rely on variables such as name, age, and state or country of birth. These challenges have contributed to relatively low rates of coverage and high rates of error and bias in linked census data from 1850-1940, especially when compared to linkages across more recent surveys and administrative data (e.g. Massey 2017; Massey et al. 2018; Bailey et al. 2020).

Linkages between twenty-first century U.S. censuses set a high standard for coverage and accuracy. With a focus on the higher-information censuses since 2000, Census Bureau researchers have 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 processes operate by matching any file to a large composite of administrative records. Valid links are assigned a unique Protected Identification Key (PIK) that facilitates linkage to any other file that has been assigned PIKs. Research based on linked census data from 2000 forward suggests that coverage is high and error in PIK assignment is minimal. For censuses and major surveys conducted since 2000, about 90 percent of cases have PIKs assigned, and nearly all assigned PIKs have been shown to be accurate (Mulrow et al. 2011, Layne et al. 2014).

Even with this high linkage rate, PIK coverage in the recent censuses varies significantly by key population characteristics. This is mainly because respondents who do not appear in administrative records cannot be assigned a PIK. For this reason, the PIKs assigned to recent censuses and surveys have been shown to over-represent the non-migrant population, those with proficiency in English, citizens, and other population sub-groups that are more likely to appear in federal government administrative records (Rastogi and O'Hara 2012). Linkage rates also vary by geography, and are lowest in the Southwest, and highest in the upper Midwest (Rastogi and O'Hara 2012, Bond et al. 2014, Layne et al. 2014).

For the 1850-1940 Censuses, linkage rates between successive censuses have typically been around 10 to 45 percent for native-born men, and much less for the whole population (Goeken et al., 2011; Helgertz et al. 2022; Massey et al. 2018). Our work linking the 1940 Census to modern administrative records fared somewhat better, producing PIKs for 40 percent of adults and 70 percent of children (Massey et al. 2018). Validation tests showed that about 4 percent of the assigned 1940 PIKs were likely to be in error (Massey et al. 2018).

The most directly relevant preliminary studies come from pilot projects conducted on small samples of the 1960 and 1990 censuses. In a pilot conducted with 1,500 hand-entered records from 1960, about 75 percent of records were successfully linked to administrative records (Massey 2014). That work found that the quarter of birth variable—which is available in 1960-1980 censuses but was not in 1940—significantly improved linkage results beyond what would have been achieved using just the age variable (Massey 2014). Used in conjunction with age and year of birth, this variable will likely allow for a finer-grained match to the exact date of birth variable in the administrative records composite file. A pilot project from the 1990 Census further suggests that we may expect linkage rates to increase with each census year. In that study, PIKs were assigned to a broadly representative sample of 28,000 digitized records collected as part of the 1990 Content Reinterview Survey (CRS) (Johnson et al. 2015). The 1990 CRS linkage team assigned PIKs to 84 percent of cases, including 87 percent of adults and 73 percent of children. We can expect higher rates of PIK assignment to the full 1990 Census file, because the CRS project did not use the full residential addresses that we have obtained for the 1990 Census respondents.

Based on the experiences of these pilot projects, we expect linkage rates to increase over time, from a low of 75 percent in 1960 to a high of 85 percent in 1990. We expect the rates to increase with each census year, since each census has slightly more information to support record linkage, and there are more administrative records available for the of the time-specific composite file that the census will be linked to. These expected rates derive from the pilot projects, and also fit squarely between the observed rates in 1940—when there is significantly less information in the census and the administrative records—and the twenty-first century censuses. In terms of expected error, our estimates can only draw on the completed linkages

from 1940 (where 4 percent of links were estimated to be in error) and the 2000 Census (where less than 1 percent of the linkages were estimated to be in error). We expect our measure of incorrect links (Type 1 errors) to fall between these two extremes. While no preliminary studies estimate the proportion of unlinked cases that should have been linked (Type 2 errors), we have developed methods to measure these in the DCDL data.

Methods

The DCDL project will adapt and expand well-established techniques to conduct record linkage across multiple decennial censuses. The core of the record linkage strategy will use a modified version of the Census Bureau's production record linkage processes, the Person Identification Validation System (PVS) (Layne and Wagner 2014; Alexander et al. 2017; Massey 2017; Massey et al. 2018). We will use probabilistic matching methods similar to those developed for the already-completed linkages for 1940 and 2000 through 2020, and assign a PIK to each case. Since the data available in each decennial census and in administrative records varies over time, we will adapt these methods to maximize the use of relevant variables during each census year.

The Census Bureau's PVS process assigns PIKs probabilistically, by linking census records to a composite database created from federal-agency administrative records that contains each person's unique PIK value. Each PIK is associated with an SSN or an Individual Taxpayer Identification Number (ITIN). ITINs are tax numbers used by people who do not have an SSN but are required to file federal income taxes. When a census or survey record is linked to the composite database, the record is assigned the relevant PIK from the composite. Respondent names are removed from the data and researchers are provided with a "PIK Crosswalk", which

includes the survey's record identification numbers (such as a household ID number and a person ID number) and a PIK for each case that was successfully linked to the composite. Census and survey data with PIK crosswalks are made available to researchers on Census Bureau approved research projects within the FSRDC's restricted environment for research. We will follow an adapted version of PVS to assign PIKs to the 1960-1990 censuses, and to make the resulting PIK crosswalks available through the FSRDC.

Table 1 shows the administrative data sources to be integrated and used for the composite reference file for linking the 1960-1990 censuses, as well as the variables available to support the linkage and assignment of PIKs in each data set. The two sources for these records are the Social Security Administration (SSA) and the Internal Revenue Service (IRS). The SSA manages the public old-age pension program and a disability supplement program for people of all ages. Most people born in the U.S. are registered for this program at birth and are assigned a permanent SSN (Genadek et al. 2022). The SSN is used to track eligibility for pension benefits that are primarily earned through credits as a worker or as a worker's dependent. The IRS data is derived from annual income information that most Americans with income are required to report each year in order to establish their taxes owed.

The core components of our administrative records composite include the SSA Numerical Identification (Numident) file, the SSA Master Beneficiary Record (MBR), and the IRS tax return data. The three primary administrative data sources in the composite will be linked to one another using the SSNs on each file. The SSNs in these files are expected to be high quality, since all of these data are used for essential government functions.

Table 1. Data sources used to construct the administrative records composite file

	Census Numident 1936-present	IRS 1040 forms 1969	IRS 1040 forms 1974, 1979, 1984, 1989, 1994	SSA Master Beneficiary Record 1962-present
Universe	all SSN holders	all tax filers	all tax filers and spouses	SSN holder who earned benefits, child or spouse receiving benefits
Name	x		x	
Alternate Name	x			
Date of Birth	x			
Date of Death	x			
Sex	x			
Marital Status		x	x	x
State or country of birth	x			
Parents' names	x			
Street address		x	x	
County		x	x	
SSN	x	x	x	x
Spouse's SSN			x	x
Child's SSN				x

Most of the composite's demographic information comes from the SSA Numident, which is built from applications for an SSN. The SSA Numident is a quarterly transactions file that includes all new applications as well as changes to existing SSNs. The Census Bureau has standardized this flow of data into a cumulative annual snapshot file called the "Census Numident." The Census Numident contains basic information on all SSN-holders going back to the beginning of the program in 1936, also including any updates the SSA received, such as new names and dates of death (a critical variable for the SSA, since the date of death usually means the suspension of pension benefits) (Finlay and Genadek, 2021). The variables that we include in the composite include SSN, name, alternate/married name, sex, date of birth, state or country of birth, parents' names and—to exclude ineligible candidate links—date of death.

The composite's information on addresses comes from the IRS tax returns. These data include the complete universe of IRS tax return files, obtained by the Census Bureau from the IRS shortly following each tax year. The Census Bureau holds these data for every fifth year

from 1969 through 1994, and every year from 1995 to the present. The files contain the original data transferred from the IRS and have information from the IRS 1040 files including SSN, marital status, mailing address, and—for 1974 forward—all of the same information as well as individual filers' and spouses' names. Our use of these files will focus on 1969, 1979, and 1989, because those files were created at almost exactly the same time as each census was conducted. For instance, 1969 IRS 1040 file includes addresses from tax returns that were likely filed on and before April 15, 1970, which is nearly the same time that the 1970 Census was conducted. Unlike the modern IRS 1040 data, the 1969-1994 IRS 1040 data do not include identifying information for household dependents. For 1969, there are only SSNs for the filer, or head of household, with no names provided. For 1974-1994 there are SSNs *and* names for the filer and—for joint returns—the filer's spouse. None of these files include information for dependents or children in the filers' households.

The third file we will use to construct the administrative records composite is the SSA's MBR. The MBR is particularly useful for specifying family relationships. The MBR provides information on Social Security benefit recipients. Since benefits are often provided to the minor child of the earner (in the case of disability or death) or the spouse of the earner (in the case of death), the MBR includes the SSN of the person who originally earned the benefit as well as the SSN of the person who received the benefit (if a different person). Since the earner and beneficiary are usually related, this file often effectively provides an SSN-to-SSN link between two spouses or between a parent and child. These relationships are explicitly indicated by a separate code on the record. These data provide an additional record connecting potential couples within a household or parent and children pairs in a household. The MBR contains information

on individuals who received benefits since 1962, which will reveal family relationships for many older respondents in the 1960-1990 censuses.

As a single integrated file, the composite will contain all the information available for each person, as well as additional constructed variables to make the information comparable to each year's decennial census data. For example, the composite's street address and county variables are converted to census geographies, specifically census tract, for each of the censuses using a variety of methods. The composite's date of birth variable will be converted to year of birth and quarter of birth, again to maximize comparability with the census data which does not include exact birth date. The SSN information in the MBR will be enriched with individual names from the Census Numident, including names for benefit recipients, as well as their spouses and parents who are specified MBR. Finally, the variables will be harmonized so all data have the same codes and formats to facilitate the linkage. The resulting file will contain one row for each person ever observed in the administrative records, with all available information about that person and any observed family members included as well.

Linking the 1960-1990 censuses to the administrative records composite.

Table 2 shows the linkage variables that can be used to probabilistically link census records to the administrative records composite file. The variables available vary only slightly by census year. For all records across all years, the linkage to the composite file will primarily rely on name, age, sex, and place of residence. For children living with a parent, parents' names will be used to link to the composite. For couples, the marital status variable and the spouse's name will be used for the linkage. In all years other than 1990, the use of quarter of birth will be investigated for matching, though birthdays are often misreported in survey data (Larsen et al.

2019). The 1990 Census contains full addresses for most respondents. In the 1960-1980, census geographies (such as census tract) and county information will be used to aid in the individual level linkage to the composite.

For each census, about one-sixth to one-fourth of households were asked to complete additional questions via the long form survey. For the subset of census respondents who received the long form questionnaire, the linkage to the composite will also use the state or country of birth variable. This variable is also available in the Census Numident file, so it is a useful variable for linking the census respondents to the composite file, even though it will only be available for some cases. Since this variable provides an additional linkage key, we will expect slightly more accuracy in the linkage of long form census records to the composite.

Table 2. Linkage variables in 1960-1990 censuses and the administrative records composite file

	Decennial Censuses Short Forms 1970-1990	Decennial Censuses Long Forms 1960-1990	Administrative Records Composite
Universe	all U.S. residents	16-25% sample of U.S. residents	Varies by field
Name	x	x	x
Age/Year of Birth	x	x	x
Quarter of birth		1960-1980	x
Sex	x	x	x
Marital Status	x	x	x
State or country of birth		x	x
Parents' names	x*	x*	x
Street address	1990	1990	x
Census Tract	x	x	x
County	x	x	x
SSN/PIK			x

*available for respondents living with parents

The 1960-1990 censuses have a large number of linkage variables, and it may seem that we could directly link one census to another without the use of an administrative records

composite. With censuses that are 10 years apart, however, the power of the composite made from administrative data is that it includes year-specific versions of information that often changes over time. For instance, the place of residence variables available in each census will be linked using the composite's time-specific location information from the IRS 1040 data. Without this information, the censuses' geographic variables would only be useful for linking those who had not moved within a 10-year period. Similarly, the composite contains all name variants that were ever reported to the SSA, including names that were changed upon marriage or for other reasons. Without the composite, the censuses' name variables would only be useful for those who had never changed their names.

Increasing PIK assignment in the 1960-1990 censuses

As described above, the linkage strategy consists of linking one file (the census) to an integrated administrative records composite, using well-established probabilistic matching techniques with low-tolerance differences between variables in the two files. These techniques produce robust linkages that have already been used to produce research in top-tier journals across the social sciences, as well as in production record linkage applications within the Census Bureau (e.g., Alexander et al. 2017; Song et al. 2020; Ferrie et al. 2021; Lowrey et al. 2021; Miller et al. 2021; Stevenson et al. 2021). For a large majority of the population—specifically those represented with high-quality data in administrative records as SSN-holders or tax filers—this approach will continue to work well, as it has in censuses before and after this period. However, to link members of the U.S. population that are not in the administrative data and are often groups researchers want to study, we will need to use new techniques.

The first approach for broader coverage will take advantage of known family links from available data. For instance, consider a married couple in the tax data where we link one spouse

to the census but not the other. In this case, we will make another pass at the census family, considering the unlinked spouse from the tax data as a candidate link for the unlinked spouse in the census. We will carry out similar processes with other available administrative data where we observe spousal links, child-parent links, and sibling links, and data such as censuses preceding and following the census being linked and proximate household survey data.

The second approach to improve coverage in the linkage of DCDL records will use new techniques for entity resolution to assign PIKs to respondents who were not initially linked to the administrative records composite. About 10 percent of respondents in 2000 were not assigned PIKs, and we expect 15 to 25 percent of census respondents in the earlier census years not to have PIKs assigned. Considering only unlinked individuals from the censuses, as well as those from administrative records who were never linked, we will look for links over time and between censuses and the other records. Rather than a file-to-file linkage, we will approach these sets as a large deduplication problem (e.g. Steorts 2015, Steorts et al. 2016). When we observe duplicate records in this set of data, we will create a new, unique PIK value that is not associated with an SSN or ITIN.

Methods for Evaluating Outcomes

Linkage quality issues are likely to be a greater challenge in the 1960-1990 census data than in the more recent decennial census data, since linkages from these years will be made with less precise linkage variables. For example, the 1960-1990 censuses do not include exact date of birth. Since the administrative records composite contains full date of birth, we need to convert (and reduce) this information into variables for age (at the time of the decennial census), year of birth, and quarter of birth. By contrast, the censuses of 2000-2020 contained exact date of birth

and are thus able to take advantage of the full birthdate information from the composite. The PIK assignment for 1960-1990 censuses will also likely rely on less precise name data, since names recovered using OCR in 1960-1990 may be lower-quality than the names recovered using OCR with clerical edits in 2000-2020 (Alexander et al. 2021; U.S. Census Bureau 2003, 2009a, 2009b, 2017). Thus, the PIKs assigned to the 1960-1990 censuses are likely to cover fewer records and have more errors than those from the 2000-2020 censuses.

We will carry out two strategies for documenting linkage quality. Our first strategy will be to replicate each year's linkage approach with the 2000 Census data, and then compare PIKs generated with the full information available in the 2000 Census to PIKs generated with the more limited information that was available in the 1960-1990 censuses. This is based on the strategy that we used to assess the quality of PIKs assigned to the 1940 Census (Massey et al. 2018). While the 2000 Census PIKs were assigned probabilistically and are therefore not real "truth data," they were assigned with more information than the 1960-1990 PIKs will be. This comparison will help us understand the cost of making these lower-information linkages, and to quantify differences in coverage and error between the 2000 Census PIK assignment and the pre-2000 PIK assignment. It will also allow us to specify links made in error (compared to the higher information data) and links missed.

For example, to estimate the quality of the 1960 links, we will conduct a new linkage with the 2000 Census respondents. We will begin by removing the exact date of birth from the 2000 data, replacing it with the level of information available in 1960 (age, year of birth, and quarter of birth). If our analysis of 1960's OCR-based name data shows that those names were of lower quality than name data from 2000 (when compared to the hand-entered "truth data" from each year), we will degrade the quality of the 2000 Census names accordingly (e.g., Kasewa et

al. 2018). Finally, we will create an administrative records composite consisting only of records from the SSA data, since no tax data are available in 1960. Using this modified base of 2000 Census data and administrative records, we will carry out the record linkage processes we used for the 1960 Census. We will then compare the "limited information" links made with the 2000 Census data to the "full information" links we had originally made for the 2000 Census. We will use similar assessments for all years to generate new information on linkage quality in the lower-information methods we use for the older censuses.

Our second strategy for documenting linkage quality will be based on data from the Panel Study of Income Dynamics (PSID). The PSID is the longest-running longitudinal household survey in the world and has been conducted at the University of Michigan's Institute for Social Research since its inception. The survey's leadership team is in the process bringing these data into the Census Bureau's Data Linkage Infrastructure. Initiated with 18,000 individuals in 5,000 families in 1968, PSID respondents were interviewed every year through 1997, and then every two years after that. The PSID has collected and validated SSNs for about 5,000 respondents.

For PSID cases with known SSNs, we will assign PIKs by linking respondents' SSNs to the administrative records composite file and re-validating each link with name and date of birth. These PIKs will present us with near-certain truth data on individuals who, in many cases, were interviewed consistently for 50 years. Using the PSID interviews that are most proximate to the 1970-2010 censuses, we will remove the PII from the PSID to make the PSID have the same PII as the proximate census data. We will create "Evaluation PIKs" on the PSID by carrying out our census linkage procedures on these lower-information PSID records. When we compare a PSID respondent's truth PIK (based on the SSNs and additional PII in the PSID) to the same individual's Evaluation PIK (based on lower-information PII from PSID interviews proximate to

census years), we will be observing the quality of the 1970-1990 linkages. We will also benefit from knowing that the PIKs on PSID data should not change over time, since they were verifiably the same individual who was interviewed by the PSID in (approximately) 1970, 1980, 1990, 2000, and 2010. When we observe an Evaluation PIK on a PSID respondent changing over time, we will be observing errors in the late-twentieth century linked data.

Our investigations of linkage quality will have a feedback effect that informs our core linkage techniques. For instance, when we identify subsets of the low-information links from 1960-1980 that are subject to significant error, we will attempt to fine-tune tolerances in the linkage process, such as the string comparator scores or age differences. Our linkage experiments using the 2000 Census and the PSID will thus serve both a documentation purpose—describing the procedures we followed and the error we observed—as well as providing us with feedback and strategies to improve the record linkage processes for censuses of 1960-1990.

Conclusion

The DCDL's data linkages will complete the largest, longest, and most substantively comprehensive longitudinal data resource in the U.S., effectively providing a link between public and confidential records over the past 170 years. We described the protocol that we will follow to link the 1960 through 1990 censuses, as well as the methods we will use to assess the quality of the linkages. Our methods make use of a key difference between census data and administrative records. The censuses are rich with content, but they have less comprehensive PII that sometimes changes between censuses. Routine administrative records, on the other hand, are more narrow in content but have the high-quality, time-specific PII that is necessary to administer social programs. In order to compensate for the fact that we are linking censuses with

limited PII that changes over time, we leverage the high-quality PII in proximate administrative records. We will use additional non-routine survey data—in the form of the Panel Study of Income Dynamics—to evaluate the quality and coverage of our links. The resulting infrastructure will allow researchers to investigate long-term demographic and socio-economic dynamics on a scale that has never before been possible.

References

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, Santiago Pérez and Myera Rashid. C*ensus Linking Project: Version 2.0* [dataset]. 2020. https://censuslinkingproject.org

Alexander, J. T., J. Fisher, and K. Genadek (2022). Digitizing Hand-Written Data with Automated Methods: A Pilot Project Using the 1990 U.S. Census. *Journal of Economic and Social Measurement* 46 (2): 95-108.

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

Antonie L, Inwood K, Lizotte DJ, Andrew Ross J. (2014). Tracking people over time in 19th century Canada for longitudinal analysis. *Machine Learning*. 95(1):129–46.

Bailey, M. J., C. Cole, M. Henderson, and C. Massey. (2020). How well do automated methods perform in historical samples? Evidence from new ground truth. *Journal of Economic Literature* 58 (4):997–1044.

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*.

Ferrie J, Massey C, Rothbaum J. Do Grandparents Matter? Multigenerational Mobility in the United States, 1940–2015 (2021). *Journal of Labor Economics* 9(3):597–637.

Finlay, K., & Genadek, K. R. (2021). Measuring all-cause mortality with the Census Numident file. *American Journal of Public Health*, 111(S2): S141-S148.

Genadek, K. R., & Alexander, J. T. (2019). The Decennial Census Digitization and Linkage Project. *Census Bureau Working Paper, No ADEP-2019-01*

Genadek, K., Sanders, J., & Stevenson, A. (2022). Measuring US fertility using administrative data from the Census Bureau. *Demographic Research*, 47(2), 37-58.

Goeken, Ron, Lap Huynh , T. A. Lynch & Rebecca Vick (2011). New Methods of Census Record Linking, *Historical Methods*, 44:1, 7-14.

Helgertz, Jonas, Joseph Price, Jacob Wellington, Kelly J Thompson, Steven Ruggles & Catherine A. Fitch (2022). A new strategy for linking U.S. historical censuses: A case study for the IPUMS multigenerational longitudinal panel, *Historical Methods* 55:1, 12-29.

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.

Helgertz, Jonas, Steven Ruggles, John Robert Warren, Catherine A. Fitch, Ronald Goeken, J. David Hacker, Matt A. Nelson, Joseph P. Price, Evan Roberts, and Matthew Sobek. *IPUMS Multigenerational Longitudinal Panel: Version 1.0* [dataset]. Minneapolis, MN: IPUMS, 2020.

Kasewa, Sudhanshu, Pontus Stenetorp and Sebastian Riedel (2018). Wronging a Right: Generating Better Errors to Improve Grammatical Error Detection. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4977–4983

King, M. L., & Magnuson, D. L. (1995). Perspectives on historical US census undercounts. *Social Science History*, *19*(4): 455-466.

Larsen AF, Headey D, Masters WA. Misreporting Month of Birth: Diagnosis and Implications for Research on Nutrition and Early Childhood in Developing Countries (2019). *Demography*. 56(2):707-728.

Layne, M., and Wagner, D. (2014). The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software. *Center for Administrative Records Research and Applications Working Paper*, *1*.

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*.

Leibbrand C, Massey C, Alexander JT, Genadek KR, Tolnay S (2020). The Great Migration and Residential Segregation in American Cities during the Twentieth Century. *Soc Sci Hist*. 44(1):19–55.

Lowrey K, Van Hook J, Bachmeier J, Foster T. (2021). Leapfrogging the Melting Pot? European Immigrants' Intergenerational Mobility across the Twentieth Century. *SocScience* 8:480–512.

Massey, C. G. (2014a). *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. (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.

Miller S, Johnson N, Wherry LR. Medicaid and Mortality: New Evidence From Linked Survey and Administrative Data. (2021) *The Quarterly Journal of Economics* 136(3):1783–1829.

Miller S, Wherry LR, Mazumder B. (2021). Estimated Mortality Increases During The COVID-19 Pandemic By Socioeconomic Status, Race, And Ethnicity. Health Aff (Millwood). 40(8):1252–1260.

Mulrow, E., Mushtaq, A., Pramanik, S. and Fontes, A. (2011). Final Report: Assessment of the U.S. Census Bureau's Person Identification Validation System. NORC at the University of Chicago. Accessed Aug 2022 at https://www.norc.org/Research/Projects/Pages/census-personal-validation-system-assessment-pvs.aspx.

Polyakova M, Udalova V, Kocks G, Genadek K, Finlay K, Finkelstein AN (2021). Racial Disparities In Excess All-Cause Mortality During The Early COVID-19 Pandemic Varied Substantially Across States. *Health Affairs* 40(2):307–316.

Rastogi, S. and Amy O'Hara. (2012). 2010 Census Match Study. *2010 Census Planning Memoranda Series*, #247.

Ruggles S, Roberts E, Sarkar S, Sobek M. (2011). The North Atlantic Population Project: progress and prospects. *Historical Methods*. 44(1):1–6.

Ruggles, Steven, Catherine A. Fitch, and Evan Roberts (2018). Historical Census Record Linkage. *Annual Review of Sociology*. 44: 19–37.

Ruggles, Steven, Catherine A. Fitch, Ronald Goeken, J. David Hacker, Matt A. Nelson, Evan Roberts, Megan Schouweiler, and Matthew Sobek. (2021). *IPUMS Ancestry Full Count Data: Version 3.0* [dataset]. Minneapolis, MN: IPUMS, 2021.

Song X, Massey CG, Rolf KA, Ferrie JP, Rothbaum JL, Xie Y. (2020). Long-term decline in intergenerational mobility in the United States since the 1850s. *Proc Natl Acad Sci U S A*. 117(1):251–258.

Steorts, R. (2015). Entity Resolution using Empirically Motivated Priors, *Bayesian Analysis*, 10(4) 849–875.

Steorts, R., R. Hall, and S.E. Fienberg (2016). A Bayesian Approach to Graphical Record Linkage and Deduplication, *Journal of the American Statistical Association*, 111(516): 1660-1672.

Stevenson AJ, Genadek KR, Yeatman S, Mollborn S, Menken JA. (2021). The impact of contraceptive access on high school graduation. *Sci Adv.* 7(19):eabf6732.

U.S. Census Bureau (2003) "Census 2000 Data Capture." Census 2000 Testing, Experimentation, and Evaluation Program: Technical Report Series, No. 3.

U.S. Census Bureau (2009a) History: 2000 Census of Population and Housing (Volume 1). U.S. Government Printing Office, Washington, DC.

U.S. Census Bureau (2009b) History: 2000 Census of Population and Housing (Volume 1). U.S. Government Printing Office, Washington, DC.

U.S. Census Bureau (2017) 2020 Census Detailed Operational Plan for: 10. Paper Data Capture (PDC) Operation. U.S. Government Printing Office, Washington, DC.

U.S. Census Bureau (2021a) Census Longitudinal Infrastructure Project (CLIP). Webpage at https://www.census.gov/about/adrm/linkage/projects/clip.html. Accessed August 2022.

U.S. Census Bureau (2021b) Census Longitudinal Infrastructure Project Pilots. Webpage at https://www.census.gov/about/adrm/linkage/projects/clip/CLIP-Projects.html. Accessed August 2022.

U.S. Census Bureau (2022) Census Longitudinal Infrastructure Project Pilots. Webpage at https://www.census.gov/programs-surveys/dcdl.html. Accessed August 2022.

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).*