ADEP WORKING PAPER SERIES

The Missing Link: Data Capture Technology and the Making of a Longitudinal U.S. Census Infrastructure

Katie R. Genadek
U.S. Census Bureau

J. Trent Alexander
University of Michigan

Working Paper 2022-02

December 2022

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

 We thank Carla Medalia, Randy Becker, and Sharon Tosi Lacey for helpful feedback and David Bleckley for excellent research assistance. This work was supported in part by the National Science Foundation under the HNDS-I: Decennial Census Linkage Project, #2023639. The published version of this paper can be found here: https://ieeexplore.ieee.org/abstract/document/9844891

The Missing Link: Data Capture Technology and the Making of a Longitudinal U.S. Infrastructure

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

ADEP Working Paper 2022-02
December 2022

Abstract
Social scientists' use of linked decennial census data has grown extensively over the past two decades. For U.S. census data before 1950, a large body of linked data has been made available within the past few years. The 2000 and 2010 decennial data have been linked to one another and back to the 1940 Census. For the censuses in between these years – from 1950 to 1990 – no linked data are available yet. This paper focuses on the technological advances in data capture that have enabled this centuries-long longitudinal data infrastructure to flourish while still leaving a sizeable "missing link" from 1950 to 1990. We will describe the development of modern technology to capture handwritten data at the Census Bureau, and ongoing efforts to digitize remaining information from and create linkages between the censuses of 1950 through 1990.

Keywords: Decennial Census, Data, Technology
JEL Codes: N00, O33

Introduction

Social scientists' use of linked decennial census data has grown extensively over the past two decades, with the number of published articles using linked census data more than quadrupling between 2000 and 2020. These articles provided ground-breaking research on various social processes in the U.S. over the century including understanding migration [6], [17], [27], social mobility [16], [45], family transitions [40], immigration [29] and public policy [38], [46]. While the growing use of linked census data has benefited from the ever-increasing computational power available to researchers, it has also been made possible by significant advances in the capture, curation, and dissemination of large-scale data resources. Data capture includes the wide range of activities necessary to prepare and preserve raw census records for use by the law-makers, academics, and genealogists who need them. At different points in time, state-of-the-art census data capture has included manual tallying, mechanical aggregation, high-speed microfilm processing, the creation of soundex-based name indexes, and most recently the digitization and dissemination of complete census responses.

For U.S. census data before 1950, a large body of linked data has been made available within the past few years, as a result of long-term efforts to digitize the images and responses from those censuses (e.g., [3], [25], [11]). The 1950 Census manuscripts were released to the public on April 1, 2022 and will be fully digitized through a combination of commercial and scientific funding. For U.S. census data from 1960 to the present, which are still protected under confidentiality requirements of Title 13 of the U.S. Code, the availability of linked data is much more limited. The 2000-2020 Census data have been linked to one another and back to the 1940 Census, and all are currently available within the Census Bureau's restricted research environment. For the censuses in between these years—from 1950 to 1990—no linked data are available yet [30]. In this paper we focus on the technological advances in data capture that have enabled this centuries-long longitudinal data infrastructure to flourish while still leaving this sizeable "missing link" from 1950 to 1990.

The lack of linked census data during the mid to late twentieth century is tightly connected to the evolution of the data capture technology used to process records directly following each census. A key piece of information necessary to link these data over time—the names of respondents—was not digitally captured in census processing before 2000 and was therefore not digitized at the time of each decennial census was conducted. This has meant that

data from every census from 1790-1990 has effectively needed to be captured twice: first by the federal government to produce required census statistics, and second—often many decades later—to capture additional information required for genealogical research and the creation of longitudinal data files. For the period from 1790-1940, that "second capture" (which actually took place over most of the twentieth century) is largely complete, and the results have been leveraged to make the newly-available linked data files. For the remaining years from 1950-1990, the second capture is just beginning.

The full second capture of historical census data has enabled scholars to address questions that would not have been possible with the small samples that were previously available. For example, with coded data on the entire population, researchers have designed innovative ways to study assimilation, international migration, social and geographic mobility, and neighborhood-based segregation [4], [28], [23]. The digitization of respondent names has also allowed for innovative. research using names [2], [8], [34], in addition to facilitating the development of a robust longitudinal infrastructure that will be used and improved by generations of social scientists and genealogists.

These linked census resources have been developed not to answer any particular research question, but rather as a public and scientific good. The pre-1950 data is well-documented, suitable for a broad range of studies, and is available for free download. The digitization of the full count 1950 Census data is now in progress. The confidential data from 1960 to the present is also available to researchers, via the Federal Statistical Research Data Centers (FSRDCs). The long-term vision is for all of these files to be linked together at the individual-level and available through the FSRDCs. Especially with the addition of the late-twentieth century "missing link" of data from 1950-1990, this multi-generational resource will permit investigations that were not possible before. With a breadth and depth that was previously out of reach, scholars and policy-makers will be able to evaluate the impact of early-life and ancestral experiences—including parental economic status, environmental exposures, policy conditions, social institutions, and neighborhood characteristics—on the social, economic, and behavioral experiences of Americans over time and through generations.

Below we describe the evolution of technology related to data capture of the decennial censuses, including the initial capture for making census statistics as well as the second capture for digitizing all remaining information on the census forms before 2000. For the censuses from

2000 onward, we detail how these two passes have taken place concurrently, as both had become essential parts of the Census Bureau's data processing needs. We conclude by describing how a collaborative project between the Census Bureau and academic institutions is currently making the necessary second capture of the 1950-1990 census forms and preparing to link the newly-digitized name data over time.

Data Capture, Digitization, and Linkage: 1790-1950 Censuses

Since the first U.S. decennial census in 1790, the federal government has been capturing and tabulating information on the full U.S. population every decade. This operational goal—the need to count and provide basic information about the resident population—has been a driving force in technological change in data capture. The Census Bureau's original capture of this information involved increasingly sophisticated tallying and mechanical processes. The information from the 1790 – 1950 Censuses was of course not stored digitally, though the records were retained and have fully been digitized in recent years. In the past three decades in particular, the entirety of the 1790-1940 censuses have been digitally re-captured, mainly using manual data entry from digitized images, and 1950 soon will be digitized as well. These large-scale census digitization projects have been leveraged to create a new generation of powerful linked data resources.

Prior to 1850, the U.S. marshals responsible for conducting the census collected information by writing each head of household's name and then tallying information on the other members of the household on the same row as the household head. The information was then summed across households within a variety of geographic areas. Beginning with the 1850 Census, the unit of census collection changed from the household to the individuals situated within a household, which increased the range of data collected and started the recording of all respondent names. The Census Office continued to use tally tabulation methods through the mid-1800s. Beginning in 1890, the Census Office transferred information from the enumeration sheets to punch cards that could be processed with new electro-mechanical tabulators [41]. The Census Office (which became the Census Bureau in 1902), continued to evolve the tabulation process through 1950.

While respondent names were not used for producing official statistics, the Census Bureau has a long history of using respondent names on original census manuscripts to support the Age Search service. The Census Bureau's Age Search service, established in 1903 and still

active as of 2021, provides certified transcripts to individuals or their legal representatives of census information that can be used as evidence of birth or citizenship [5], [54]. From 1935 to 1943, the Census Bureau led a large-scale Works Progress Administration (WPA) project to create indexes of respondent names from the 1880-1930 censuses. Intended to increase the efficiency of the Age Search operations, thousands of employees handwrote respondents' name, age, and sex onto index cards that also specified each record's enumeration district and page number. Those cards were then sorted by the four-digit "Soundex" code of the household head's last name and stored on microfilm [37]. The WPA projects resulted in a full name enumeration of the 1900 and 1920 Censuses, and a partial name list for the 1880, 1890, 1910, and 1930 Censuses, totaling more than 400 million names [12], [54].

The paper forms of the pre-1950 censuses were transferred to microfilm in a similarly massive operation carried out by the Census Bureau starting in the late 1930s [18], [47]. Between 1937 and 1944, the Census Bureau microfilmed the paper forms from the 1840-1880 censuses and 1900-1940 censuses, to reduce the ongoing deterioration of the records and to save floor space [18]. The operation ultimately allowed the agency to replace miles of shelving that held bound census volumes with several dozen square feet of microfilm cabinets [18]. The 6.5 million pages from the 1950 Census population schedules were microfilmed in 1951 [1], [31]. The resulting microfilm reels were archived with the National Archives and Records Administration (NARA), and the Census Bureau retained a copy to support Age Search and other operational tasks. The Census Bureau treats all decennial census responses as confidential and protected by Title 13 U.S. Code regardless of the decennial year, yet since 1952, NARA has released the full census records 72-years following a census [36]. In 1978, this was codified in Public Law 95-416, and NARA releases the census responses publicly after 72 years. For example, on April 1, 2012, NARA publicly released all of the digital images from the 1940 Census.

Social scientists and genealogists have made significant use of the WPA-funded census indexes and the microfilmed census manuscripts. Small samples of hand-linked census records were a key source in some of the classic works of social history (e.g., [35], [48]). Beginning in the 1980s and 1990s, researchers at a number of institutions—most notably the University of Minnesota—produced nationally-representative microdata samples of the microfilmed census records by hand-keying and coding microdata from the nineteenth and early-twentieth century censuses, which were the most recent publicly-available at that time [44]. University of

Wisconsin researchers partnered with the Census Bureau to create similar samples from the then-confidential microfilmed records from the 1940 and 1950 Censuses [50]. Since the Census Bureau itself had created the first-ever public use microdata samples beginning with the 1960 Census (and continuing to the present), by the early 2000s researchers had access to a long series of cross-sectional samples from the decennial censuses.

While decadal cross-sectional microdata samples have enabled major breakthroughs in the social sciences, the available samples of fully-digitized census records are still relatively small. The public microdata samples typically contain one percent samples of the population. Most of the public microdata files from 1850-1930 include respondent names, with the exception of the 1940 and 1950 samples which do not include names due to confidentiality restrictions in Title 13 U.S. Code. Some scholars undertook the laborious task of linking large representative microdata samples to genealogical indexes of respondent names from the surrounding censuses (e.g., [19], [24]). This early work showed the power or forming a bridge between a resource made for social scientists and a resource made for genealogists, a precursor of academic-genealogical collaborations that have continued to this day.

By the early 2000s, major genealogy companies had begun digitizing images and a limited set of information from the census microfilm released by NARA, making all of this information available online for genealogical research. Far easier to use than the previously-available soundex-based indexes to the public microfilm records, these new databases permitted researchers to query massive numbers of census records and quickly view digital images that previously would have required a microfilm search. The Integrated Public Use Microdata Series (IPUMS) project at the University of Minnesota partnered with FamilySearch and Ancestry.com to complete the digitization of images and data from all of the pre-1950 censuses, and to make the resulting data publicly available to researchers without identifying information [39]. For researchers who establish a data security agreement, IPUMS also provides the full-count 1850-1940 data with respondent names, which allows for the individual-level linkage of the records over time. By the early 2020s, the digitization of censuses from 1790 to 1940 was essentially complete. In early 2022, IPUMS and Ancestry.com are currently in the process of digitizing the newly-released 1950 Census and adding it to this same infrastructure.

Data Capture in the Digital Era: 1960-2020 Censuses

Before the 1960 Census, state-of-the-art data processing technology involved using mechanical punch card tabulation to process census forms that were comprised almost entirely of handwriting. This all changed for the 1960 Census, when forms were optimized for machine-oriented digital processing from the start. The earliest use of digital-capture technology was designed so that a subset of items could be captured through "bubbles" on forms completed by enumerators, and has evolved to capture nearly all items from respondent handwriting in 2020. Leading up to 1960, the Census Bureau worked with the National Bureau of Standards (NBS), now National Institute of Standards and Technology (NIST), to design an optical sensing system that would read "bubbled" items by shining light through microfilmed images of the original paper forms. The resulting Film Optical Sensing Device for Input to Computers (FOSDIC) was created to read the microfilmed images of decennial census responses and create individual level microdata to be tabulated by computers. Since FOSDIC relied on microfilm, this meant that the Census Bureau would continue the process of filming all decennial census forms, but not just for the purpose of archiving and to support the Age Search operation, but for the purpose of data processing itself. The microfilm-based FOSDIC system was faster than using paper and having employees enter information on punch cards, and it allowed for the immediate compact storage of the images on microfilm and destroying of large amounts of paper forms.

As a part of the 1960 Census, the Census Bureau sent all households an "Advance Census Report" form, on which household members could answer the few questions that were asked of all households. Enumerators then visited every household and either interviewed the respondent or used the Advanced Census Report to complete the FOSDIC bubble form. The enumerators also dropped off a longer questionnaire to one-in-four specified sample households, and respondents were instructed to mail this form back to the Census Bureau District Office, where Census Bureau staff transcribed responses onto a FOSDIC-ready bubble form [53].

When the Census Bureau had completed the 1960 FOSDIC forms, the forms were filmed, stored on microfilm rolls, and the microfilm were then processed through the FOSDIC machine to make microdata records. All 1960 paper responses and FOSDIC forms were destroyed following filming of the images. The microdata created by the FOSDIC was then used to make aggregate data from the 1960 Census. The Census Bureau retained the 1960 microdata from the "long form" questionnaire completed by one-in-four households, but did not retain the "short form" microdata completed by the remaining households [53]. From the long form microdata, the

Census Bureau also created and disseminated the first-ever Public Use Microdata Samples (PUMS), which included anonymized full responses for a sample of individuals for researchers to use. The 1960 Census provided the first representative PUMS file from any U.S. census and provided a model for all subsequent PUMS files that have been created from 1790-2010 [42].

The Census Bureau adapted and improved the FOSDIC over the next three decades, increasing the speed of capture from 3,000 items per minute in 1960 to 70,000 items per minute in 1990 [52]. Like all previous U.S. censuses, the tabulation of data did not require the use of names, and the FOSDIC machine only captured information obtained from filled in bubbles on the form. Some variables with handwritten responses (such as the occupation and industry fields) were typed manually, as they were required for published tabulations. Respondent names were not captured at the time of the decennial census, though they were retained in their original handwritten form on the microfilmed images. Starting in 1970, the Census Bureau mailed out FOSDIC-readable forms directly to households. Respondents were asked to complete the FOSDIC form and return it the Census Bureau, where it would be filmed and run through the FOSDIC machine [61]. This reduced the need for enumerators to re-transcribe short-forms in each home, and for Census Bureau staff to re-transcribe long forms in the District Office. Thus, in 1960, respondents' names were written by enumerators, while in 1970-1990 the handwritten names were entered by individual respondents from households.

Following the 1970 Census, the Census Bureau retained all of the microdata captured from the short and long forms by the FOSDIC, and they then used the long form microdata to produce six PUMS file to support researchers. These practices continued through the 1990 Census (and continue to the present). Census Bureau employees were able to use the full 1960 long form or sample microdata and the 1970-1990 short and long form microdata for research and other production purposes, while scholarly research outside of the Census Bureau primarily used the PUMS files from these years. With the expansion of the Census Research Data Center (now the Federal Statistical Research Data Center) program in 1998, the full confidential microdata files were made available to academic researchers in this restricted environment [33].

The original microfilm reels with the images of the census forms read through the FOSDIC machine were copied, and the copy was archived with NARA following their use at the Census Bureau. The original microfilm reels with 1960, 1970, and 1980 short form images were retained at the Census Bureau for the Personal Census Records Branch operations (the current

administrative home of the Age Search Service). A copy of the film with the 1990 short and long form information was also retained for these purposes. The microfilm resides at the Census Bureau's National Processing Center in Jeffersonville, IN (see [39]).

For the 2000 Census, the Census Bureau contracted with several outside vendors to capture and digitize the responses. The paper forms were mailed out and mailed back as they had been in previous years, but the FOSDIC process was no longer used to read the bubbled forms. Instead, the forms were scanned into digital images (rather than microfilm), and the bubbled responses were processed from the digital images using Optical Mark Recognition (OMR) technology. Write-in responses (including names) were captured with a combination of Optical Character Recognition (OCR) and hand-keying [59]. Unlike any previous census, the 2000 Census form was optimized for the digital capture of text, including instructions to respondents to print the letters of each word, with a segmented space on the form for each letter. This was the first time that names were captured in digital form at the time of the decennial census, allowing for the individual-level linkage of these data to other data with name and other identifying information. Similar to previous years, the resulting microdata was saved after tabulation for use by researchers and employees at the Census Bureau. Unlike the previous years, the 2000 Census was the first year that the Census Bureau did not retain a copy of the census form images for agency use. The Census Bureau archived the images for NARA on digital tape, and put the images on microfilm at the insistence of the genealogy community and NARA at the time [60]. Following the delivery of the images to NARA, the Census Bureau deleted their copies of the images.

The 2010 Census responses were captured with similar methods as the 2000 Census, including the use of OMR for bubbled responses and a combination of OCR and keyed data for written responses. In contrast with the 2000 Census, the 2010 Census images are retained at the Census Bureau, in addition to all machine-readable information from the census. The digital copies of images from the 2010 Census forms were shared with NARA for archival purposes, and the images are also stored and used within the Census Bureau, along with the resulting microdata, through the Census Bureau's Census Image Retrieval Application (CIRA).

While much of the 2020 Census was obtained via internet response of households, the Census Bureau captured information from paper responses using its own software, the integrated Computer Assisted Data Entry (iCADE) system. Since 52% of households completed the online

self-response to the 2020 Census (and another 34% were captured via non-response follow up), most of the 2020 results are available only as data files and not as images [56], [49]. The Census Bureau did store the images in CIRA for the small subset of the population who responded via paper (totaling about 12% of housing units).

Digitizing Respondent Names and Linking Records in the Modern Censuses

For the pre-1950 Censuses, respondent names were digitized in a "second pass" over the past 30 years, leveraging the work of the WPA, the census microfilming project, and commercial efforts largely funded by genealogical organizations. The recently-released 1950 Census is currently being digitized in this same way. The 1960-1990 Census records are still protected by Title 13 of U.S. Code, and the names were not digitized at the time of the censuses due to cost. The Census Bureau estimated that manual entry of names from the 1950 census would have cost $6 million in 1950 (or near $70 million in today's dollars) [12]. This changed with the 2000 Census, when the use of OCR allowed the agency to capture names to facilitate deduplication in responses [59]. Thus, there is a "missing link", resulting in the 1950-1990 Censuses unable to be linked over time because the names are not digitized and captured.

The Census Bureau did capture subsets of names in various exercises related to the decennial censuses of 1960 to 1990. For all of these censuses, follow up surveys were conducted shortly after the main enumeration to assess the quality of the responses collected [51]. Starting with the 1960 Census, many of the follow up surveys included linking Current Population Survey (CPS) respondents to the decennial census. The 1980 and 1990 follow up surveys also included the collection and manual keying of names for a sample of decennial census respondents. In 1980, two CPS months (April and August) were used as follow up samples, and an additional sample of 110,000 households was taken directly from the 1980 Census and resurveyed [62]. The 1990 Content Reinterview Survey followed up via telephone to reinterview 12,800 households. In addition to the 1990 Content Reinterview Survey, head of household surnames were keyed for about 4.7 million households as a part of the 1990 Census Non-Response Follow Up (NRFU) operation [63]. The 1990 Census data were also used as the sample frame for the 1993 National Survey of College Graduates (NSCG). The Census Bureau keyed a sample of 208,393 names directly from the 1990 microfilm for the NSCG sample in 1993 [10].

In addition to the hand-keying of names for these quality assurance projects, there were various efforts to advance the machine-capture of handwriting at the Census Bureau with the hope of capturing all the names in order to use them for deduplication. The Census Bureau explored the use of automation for the 1990 Census [9]. While these methods were not used in 1990, the research continued. The Census Bureau partnered with NIST to host an experiment and conference to identify the leading edge OCR technologies with the intent on using these technologies to capture handwriting in the 2000 Decennial Census. The first conference took place in May 1992 and focused on printed characters, while the second one in February 1994 included many of the same research teams but focused on more realistic capture of handwritten words from census forms and from microfilm images [64], [21]. The conclusion from the conferences was that "machine performance in reading words and phrases may now be good enough to decrease the cost and time needed to carry out a Census without decreasing the accuracy of the results." [21]

Respondent names from the 2000 Census were captured, in part, using OCR, with some names hand-keyed. This presented the first opportunity for the Census Bureau to capture the names of all respondents at the time of initial data capture without the expense of labor to hand enter the names. The names were obtained for the purposes of deduplication at the individual-level and eventually used for this purpose [59]. The names were also used for the editing and imputation of questions on sex and Hispanic origin. The names also allowed the Census Bureau to assign unique linkage keys to the data, making it the first decennial census that could be linked at the individual level to other censuses and surveys.

The data capture system used for the 2000 Census was developed and implemented by private contractors, though the Census Bureau concurrently developed an internal automated data capture system, first used in the 2002 Economic Census. The iCADE system was designed to process digital images of paper respondent questionnaires through OMR, OCR, and Key From Image (KFI) technology. The system has since been expanded to capture the American Community Survey (ACS), and for all economic surveys and economic censuses since 2000. For the 2010 Census, iCADE was used for the special censuses. The remaining part of the 2010 Census capture was also done by a contractor, including a combination of OCR and hand-keying of names. By the 2020 Census, all paper decennial census forms (16 variations of them) were captured using iCADE, including the automated capture of names.

Filling in the Missing Link: Digitizing Names From 1950-1990

Genealogists and social scientists have long sought a comprehensive, high-quality, searchable transcription of all census records, including all respondents and variables. The work of building this infrastructure has taken place gradually over the past century, requiring technological innovations and funding from government, business, and the sciences. Using new technology for storing and indexing census images in the 1930s and 1940s, the Census Bureau built a massive microfilm archive that was leveraged, improved, and ultimately digitized in the late-twentieth century. Fully-digitized census records and images are now largely available for the 1850 through 1940 Censuses, including well-documented and pre-linked data files ready for research. Likewise, the Census Bureau utilized new data capture technology to capture the names of census respondents starting for the 2000 Census, starting the Census Bureau's Linkage Data Infrastructure that is available to scholars through the FSRDCs. We have argued that these extraordinary linked data resources are the product of continual advances in data capture, data processing, data dissemination, and data storage technology, along with generations of strategy and effort from government employees, genealogical groups, and academic infrastructure-builders.

Like the censuses before and after, the 1950 through 1990 Census forms included respondents' handwritten names, even though they were not keyed at the time of census processing. While the paper forms were destroyed, the images were stored on microfilm for processing and archiving. The 1950 Census manuscripts were released to the public on April 1, 2022 and will be fully digitized through a combination of commercial and scientific funding. For 1960 through 1990, we have secured a combination of scientific funding and Census Bureau support to digitize the decennial census images and respondent names, both for the purposes of supporting Age Search and for completing the longitudinal infrastructure that has been a centuries-long American project [22].

The digitization of names from the 1960-1990 Censuses presents several unique challenges. Unlike the most recent censuses, the forms from these periods were not optimized for name capture. For instance, the instructions do not state that handwriting should be printed, and there are not individual boxes for each letter as we see in more recent surveys and censuses. As still-confidential records, names from the 1960-1990 Censuses need to be digitized in a secure computing environment, with all work done by Census Bureau employees who meet stringent

security requirements. Finally, the scale of the project is uniquely massive: in the 2020 Census, there were about 20 million names captured from handwriting. From all of the 1790-1940 censuses combined, there were about 680 million respondent names, all of which written by marshals or census enumerators and digitized by several groups over decades [58]. In the 1960-1990 Censuses, over 850 million respondent names need to be recovered, all provided in the respondents' own handwriting, often in cursive.

A project of this scale is now possible mainly through a series of technological and methodological innovations. The first is the advancement of the use of the OCR to capture handwriting generally [15], [20], [26], [43], and more specifically capturing handwriting from historical forms [13], [14]. Leveraging these developments in machine learning and OCR science, it is possible to capture the handwritten names from census forms with the accuracy necessary for linkage purposes, though still with less accuracy than is typical for hand-keyed data [7]. The general availability and decreasing costs of high-powered computers to perform OCR—as well as the storage necessary to support the images—is driving the advancements in OCR science [32], and it also makes the capture of names from 1960-1990 censuses feasible. The digitized images necessary for OCR are estimated to take up 4 petabytes (equivalent to 4,000 terabytes or 4,000,000 gigabytes) of storage space. A decade ago, this was a show-stopping storage need even for an agency the size of the Census Bureau. In 2022, it is considered a very large but feasible amount of storage space. Finally, the computing power necessary to capture the names from hundreds of millions of high-quality images would have been nearly impossible without the efficient scalability of modern cloud-based computing infrastructures.

The Decennial Census Digitization and Linkage (DCDL) project, began the work of digitizing, capturing, and linking the "missing link" in 2021. This large project will result in the creation of the longest individual-level panel of census data in the U.S., filling in the gap in the Census Bureau's Data Linkage Infrastructure [57], and disseminating these data for all scholars through the FSRDC. The research possibilities using these data are great and are only increased with the linkage of administrative and survey data to this backbone of linked census data from 1850 through the present. This project would not be possible without the long history of technological advancements in data capture and storage of U.S. census data.

It is possible the researchers in several decades or centuries will recreate much of this data to higher standards, using even better handwriting recognition, more efficient data entry, or

referencing information against other proximate data resources. If the past is any indication, we can expect it. We have seen scholars and genealogists alike progress through the laborious searching of paper forms, soundex-based searches of paper and microfilm records, computer searches of higher-quality genealogical name indexes, and now the computer-assisted search, automated handwriting capture, and linkage of full-count census data. Our successors will surely develop even better techniques to capture, link, and measure error in all of these records, to the benefit of genealogists, scientists, and the government agencies who have an interest in high-quality linked data.

References

1. "1950 Population Schedules now being microfilmed," Personnel Standards and Develop. Branch, Personnel Division, U.S. Census Bur., Washington, DC, USA, Census Bulletin, vol. I, no. 29, 1951.

2. R. Abramitzky, L. Boustan, and K. Eriksson, "Do immigrants assimilate more slowly today than in the past?," *Amer. Econ. Rev.: Insights,* vol. 2, no. 1, pp. 125–141, 2020, doi: 10.1257/aeri.20190079

3. R. Abramitzky, L. Boustan, and M. Rashid. "Census Linking Project: V1.0" [dataset]. (2020). https:// censuslinkingproject.org

4. R. Abramitzky, L. P. Boustan, and K. Eriksson, "Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration," *Amer. Econ. Rev.,* vol. 102, no. 5, pp. 1832–1856, 2012, doi: 10.1257/aer.102.5.1832.

5. "Age Search 75 years young." Personnel Standards and Develop. Branch, Personnel Division, U.S. Census Bur., Washington, DC, USA, Census Bulletin, vol. 28, no. 7, 1978.

6. J. T. Alexander, C. Leibbrand, C. Massey, and Stewart Tolnay, "Second-generation outcomes of the great migration," Demography, vol. 54, no. 6, pp. 2249-2271, Sep. 2016.

7. J. T. Alexander, J. D. Fisher, and Katie R. Genadek, "Digitizing hand-written data with automated methods: A pilot project using the 1990 U.S. Census," Associate Directorate for Econ. Programs, U.S. Census Bur., Washington, DC, USA, Census Bur. Working Paper Series ADEP-WP-21-06, Dec. 2021.

8. S. Bazzi, M. Fiszbein, and M. Gebresilasse, "Frontier culture: The roots and persistence of 'rugged individualism' in the United States," *Econometrica,* vol. 88, no. 6, pp. 2329–2368, 2020, doi: 10.3982/ECTA16484

9. P. Bounpane, "How increased automation will improve the 1990 census of population and housing of the United States," J. of Official Statist., vol. 2, no. 4, pp. 545-553, Dec. 1986.

10. C. Briseno, L. Giesbrecht, and P. McGuire, "Overview: 1993 National Survey of College Graduates," unpublished.

11. K. Buckles and J. Price, "Census tree links." Jul. 13, 2021. Distributed by Inter-university Consortium for Political and Social Research. doi: 10.3886/E144904V2.

12. "Business as usual," Personnel Standards and Develop. Branch, Personnel Division, U.S. Census Bur., Washington, DC, USA, Census Bulletin, vol. 8, no. 18, 1958.

13. H. Cao, K. Subramanian, X. Peng, J. Chen, R. Prasad, and P. Natarajan, "Extracting information from handwritten content in census forms," Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), 2012, pp. 306-309.

14. H. Cao, S. Rawls, and P. Natarajan, "1990 US Census Form Recognition Using CTC Network, WFST Language Model, and Surname Correction," 2017 14th IAPR Int. Conf. on Document Anal. and Recognit. (ICDAR), 2017, pp. 977-982, doi: 10.1109/ICDAR.2017.163.

15. V. Carbune et al., "Fast multi-language LSTM-based online handwriting recognition," Int. J. on Document Anal. and Recognit. (IJDAR), vol. 23, no. 2, pp. 89–102, Feb. 2020, doi: 10.1007/s10032-020-00350-4.

16. R. Chetty, N. Hendren, M. R. Jones, and S. R. Porter, "Race and Economic Opportunity in the United States: An Intergenerational Perspective," Quart. J. Econ., vol. 135, no. 2, Dec. 2019, doi: 10.1093/qje/qjz042.

17. W. J. Collins and M. H. Wanamaker, "Selection and economic gains in the Great Migration of African Americans: New evidence from linked census data," Amer. Econ. J.: Appl. Econ., vol. 6, no. 1, pp. 220–252, Jan. 2014, doi: 10.1257/app.6.1.220.

18. L. H. Dreiser, "Microfilm save expensive storage," Domestic Commerce, vol. 32, no. 8, pp. 13 and 24-25, 1944.

19. J. P. Ferrie, "A new sample of males linked from the public use microdata sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedules," Historical Methods: J. Quantitative and Interdisciplinary Hist., vol. 29, no. 4, pp. 141–156, Oct. 1996, doi: 10.1080/01615440.1996.10112735.

20. B. Finkelstein and K. Kuncan, "The study of handwriting recognition algorithms based on neural networks," Int. J. Hybrid Inf. Technol., vol. 14, no. 1, pp. 69–80, Mar. 2021, doi: 10.21742/ijhit.2021.14.1.05.

21. J. Geist et al., "The Second Census Optical Character Recognition Systems Conference," Nat. Inst. of Standards and Tech., U.S. Dept. of Commerce, Gaithersburg, MD, USA, NISTIR 5452, May 1994.

22. K. R. Genadek, and J. T. Alexander, "The Decennial Census Digitization and Linkage Project," Associate Directorate for Econ. Programs, U.S. Census Bur., Washington, DC, USA, Census Bur. Working Paper Series ADEP-WP-19-01, Dec. 2019.

23. A. Grigoryeva and M. Ruef, "The historical demography of racial segregation," *Amer. Sociol. Rev.*, vol. 80, no. 4, pp. 814–842, 2015, doi: 10.1177/0003122415589170

24. A. M. Guest, "Notes from the National Panel Study: Linkage and migration in the late nineteenth century", Historical Methods: J. Quantitative and Interdisciplinary Hist., vol. 20, no. 2, pp. 63-77, Apr. 1987, doi: 10.1080/ 01615440.1987.9955260.

25. J. Helgertz et al., "IPUMS multigenerational longitudinal panel. V1.0." 2020. Distributed by IPUMS. doi: 10.18128/D016.V1.0.

26. D. Keysers, T. Deselaers, H. A. Rowley, L. L. Wang, and V. Carbune, "Multi-language online handwriting recognition," IEEE Trans. on Pattern Anal. and Mach. Intell., vol. 39, no. 6, pp. 1180–1194, Jun. 2017, doi: 10.1109/tpami.2016.2572693.

27. C. Leibbrand, C. Massey, J. T. Alexander, and S. Tolnay, "Neighborhood attainment outcomes for children of the great migration," Amer. J. Sociol., vol. 125, no. 1, pp. 141–183, Jul. 2019, doi: 10.1086/703682.

28. T. D. Logan and J. M. Parman, "The national rise in residential segregation," *J. Econ. Hist.*, vol. 77, no. 1, pp. 127–170, 2017, doi: 10.1177/0003122415589170.

29. K. Lowrey, J. Van Hook, J. D. Bachmeier, and T. B. Foster, "Leapfrogging the melting pot? European immigrants' intergenerational mobility across the twentieth century," Sociological Sci., vol. 8, pp. 480-512, 2021.

30. C. G. Massey, K. R. Genadek, J. T. Alexander, T. K. Gardner, and A. O'Hara. 2018. "Linking the 1940 US Census with modern data." Historical Methods: J. of Quantitative and Interdisciplinary Hist., vol. 51, no. 4, pp. 246–257, Oct. 2018, doi: 10.1080/01615440.2018.1507772.

31. "Microfilming of P-1 Schedules complete," Personnel Standards and Develop. Branch, Personnel Division, U.S. Census Bur., Washington, DC, USA, Census Bulletin, vol. II, no. 14, 1952.

32. O. Mujtaba Khandy and S. Dadvandipour, "Analysis of machine learning algorithms for character recognition: A case study on handwritten digit recognition," Indonesian J. Elect. Eng. and Comp. Sci., vol. 21, no. 1, pp. 574-581, Jan. 2021, doi: 10.11591/ijeecs.v21.i1.pp574-581.

33. "Notice of program and request for proposals," U.S. Census Bur.,Washington, DC, USA, Federal Register, vol. 63, no. 14, 1998 https://www.govinfo.gov/content/pkg/FR-1998-01-22/ pdf/98-1657.pdf (accessed Nov. 21, 2021).

34. C. Olivetti and M. D. Paserman, "In the name of the son (and the daughter): Intergenerational mobility in the United States, 1850 –1940," *Amer. Econ. Rev.*, vol. 105, no. 8, pp. 2695–2724, 2015, doi: 10.1257/aer.20130821.

35. D. H. Parkerson, "How mobile were nineteenth-century Americans?" Historical Methods: J. Quantitative and Interdisciplinary Hist., vol. 15, no. 3, pp. 99–109, Jul. 1982, doi: 10.1080/01615440.1982.10594085.

36. R. V. Peel, "Letter to Wayne C. Grover, Archivist of the United States." Aug. 26, 1952. https://www.census.gov/ history/pdf/grover-8-26-1952.pdf (accessed Nov. 21, 2021).

37. C. Prechtel-Kluskens, "The WPA Census Soundexing projects," Genealogy Notes, vol. 34, no. 1, spring 2002, https://www.archives.gov/publications/prologue/2002/spring/soundex-projects.html (accessed Nov. 22, 2021).

38. E. Rauscher, “Does educational equality increase mobility? Exploiting nineteenth-century U.S. compulsory schooling laws,” Amer. J. Sociol., vol. 121, no. 6, pp. 1697–1761, May 2016, doi: 10.1086/685443.

39. S. Ruggles, “Collaboration of genealogy and social science history: The case of IPUMS,” Nov. 2019, https://users.pop.umn.edu/~ruggles/Articles/IPUMS_Genealogy.pdf (accessed Nov. 21, 2021).

40. S. Ruggles, “Intergenerational coresidence and family transitions in the United States, 1850–1880.” J. Marriage and Family, vol. 73, no. 1, pp. 136–148, Jan. 2011, doi: 10.1111/j.1741-3737.2010.00794.x.

41. S. Ruggles and D. L. Magnuson, “Census technology, politics, and institutional change, 1790–2020.” J. Amer. Hist., vol. 107, no. 1, pp. 19–51, Jun. 2020, doi: 10.1093/jahist/jaaa007.

42. S. Ruggles, M. Schroeder, N. Rivers, J. T. Alexander, and T. K. Gardner, “Frozen film and FOSDIC forms: Restoring the 1960 U.S. Census of Population and Housing,” Historical Methods: J. Quantitative and Interdisciplinary Hist., vol. 44, no. 2, pp. 69–78, Apr. 2011, doi: 10.1080/ 01615440.2011.561778.

43. M. A. Shubh, F. Hassan, G. Pandey, and S. Ghosh, “Handwriting recognition using deep learning,” in Emerging Trends in Data Driven Computing and Communications: Proceedings of DDCIoT 2021, R. Mathur, C. P. Gupta, V. Katewa, D. S.. Jat, N. Yadav, Eds., Singapore, Springer Nature Singapore, 2001, pp. 67-81.

44. M. Sobek and S. Ruggles. “The IPUMS project: An update,” Historical Methods: J. Quantitative and Interdisciplinary Hist., vol. 32, no. 3, pp. 102–110, Jan. 1999, doi: 10.1080/ 01615449909598930.

45. X. Song, C. G. Massey, K. A. Rolf, J. P. Ferrie, J. L. Rothbaum, and Y. Xie, “Long-term decline in intergenerational mobility in the United States since the 1850s,” Proc. Nat. Acad. Sci., vol. 117, no. 1, pp. 251–258, Nov. 2019, doi: 10.1073/pnas.1905094116.

46. A. J. Stevenson, K. R. Genadek, S. Yeatman, S. Mollborn, and J. A. Menken. “The impact of contraceptive access on high school graduation.” Science Advances, vol. 7, no. 19, eabf6732, May 2021.

47. L. D. Szucs and M. Wright, Finding Answers in US Census Records. Orem, UT, USA: Ancestry Publishing, 2001.

48. S. Thernstrom, Poverty and Progress. Cambridge, MA, USA: Harvard University Press, 1964.

49. “Updates to the 2020 Archiving Operation Detailed Operational Plan,” U.S. Census Bur., Washington, DC, USA, 2020 Census Program Memorandum Series: 2021.13, Jun. 10, 2021, https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/plan/memo-series/2020-memo-2021_13.html (accessed Nov. 21, 2021).

50. U.S. Bureau of the Census, “Census of Population, 1940 [United States]: Public Use Microdata Sample.” 1984. Distributed by Inter-university Consortium for Political and Social Research. doi: 10.3886/ICPSR08236.v1.

51. U.S. Bureau of the Census, The 1950 Census – How They Were Taken, Publications Distribution Section, U.S. Bureau the Census, Washington, DC, USA, 1955.

52. U.S. Census Bureau, “1960 Overview,” https:// www.census.gov/history/www/through_the_decades/ overview/1960.html (accessed Nov. 21, 2021).

53. U.S. Census Bureau, 1960 U.S. Censuses of Population and Housing: Procedural History. Washington, DC, USA: GPO, 1966.

54. U.S. Census Bureau, Age Search Information. Washington, DC, USA: GPO, 2000.

55. U.S. Census Bureau, “Age Search Service,” Nov. 19, 2021, https://www.census.gov/topics/population/genealogy/agesearch.html (accessed Nov. 21, 2021).

56. U.S. Census Bureau. April 26, 2021. 2020 Census Operation Quality Metrics: Release 1. Accessed on November 21, 2021 at https://www.census.gov/newsroom/press-releases/2021/ quality-indicators-on-2020-census.html.

57. U.S. Census Bureau, “Data Linkage Infrastructure,” Oct. 8, 2021, https://www.census.gov/about/adrm/linkage.html (accessed Nov. 21, 2021).

58. U.S. Census Bureau, Historical Statistics of the United States, Colonial Times to 1970, Part I. Washington, DC, USA: GPO, 1975.

59. U.S. Census Bureau, History: 2000 Census of Population and Housing. (Volume 1). Washington, DC, USA: GPO, December 2009.

60. U.S. Census Bureau, History: 2000 Census of Population and Housing. (Volume 2). Washington, DC, USA: GPO, December 2009.

61. U.S. Census Bureau, U.S. Censuses of Population and Housing 1970: Procedural History PCH-1. Washington, DC, USA: GPO, 1976.

62. U.S. Census Bureau, U.S. Censuses of Population and Housing 1980: Procedural History. Washington, DC, USA: GPO, 1989.

63. U.S. Census Bureau, U.S. Censuses of Population and Housing 1990: Procedural History. Washington, DC, USA: GPO, 1996.

64. R. A. Wilkinson et al., 1992. “The First Census Optical Character Recognition Systems Conference,” Nat. Inst. of Standards and Tech., U.S. Dept. of Commerce, Gaithersburg, MD, USA, NISTIR 4912, Aug. 1992.