IPUMS Multigenerational Longitudinal Panel

DRAFT

Steven Ruggles, Catherine Fitch, Ron Goeken, J. David Hacker, Jonas Helgertz,
Evan Roberts, Matt Sobek, Kelly Thompson, John Robert Warren, and Jacob Wellington

May 15, 2019

IPUMS
Institute for Social Research and Data Innovation
University of Minnesota

Background and Significance

IPUMS-MLP will consist of nine censuses covering the entire U.S. population enumerated between 1850 and 1940 linked with public historical administrative data from Social Security, the military, and vital registration. The linked database will be invaluable for analyzing the impact of early life conditions on health and well-being in later life, and the large scale of the resource will allow study of very small population subgroups. IPUMS-MLP is not designed to answer any particular scientific question. Rather, we plan general-purpose data infrastructure, a permanent resource that can be continuously expanded to incorporate the latest data sources as they become available, ensuring its usage for decades to come.

Former Census Bureau Director Robert Groves drew an insightful distinction between “designed data” and “organic data” [1]. Designed data, such as censuses and surveys, are created entirely to obtain information. Organic data are byproducts of transactions, including administrative records generated by Social Security, Medicare, the Internal Revenue Service, and the Armed Forces. Research on population aging currently relies primarily on designed data, despite the enormous potential of organic data to enrich our analyses. Groves argued that “the biggest payoff will lie in new combinations of designed data and organic data, not in one type alone.” Used in isolation, organic data have profound limitations that reduce their usefulness. They tend to be voluminous but shallow; they often are unrepresentative of the general population; and they frequently omit basic information about demographic behavior, economic status, education, work, and living conditions. IPUMS-MLP will enrich large sources of organic data—including Social Security, Medicare, and military records—by linking them to a century of designed census and survey data, thereby overcoming limitations of the organic data sources.

Linking individuals from childhood to old age and death through both designed and organic data allows study of aging as a process over the entire life course, not just over a few years. Indeed, IPUMS-MLP will enable investigators to extend longitudinal analysis beyond individual

life histories to investigate and understand processes of change over multiple generations [2]. In his 2010 presidential address to the Population Association of America, Robert Mare [3] argued that "the study of intergenerational mobility and most population research are governed by a two-generation (parent-to-offspring) view of intergenerational influence, to the neglect of the effects of grandparents and other ancestors and nonresident contemporary kin." Mare called for the development of sources and methods that will support analysis of change over multiple generations. IPUMS-MLP will meet this need, allowing investigators to trace records back across multiple generations and making it possible for the first time to study the transmission of characteristics and behavior across centuries.

Limitations of Current Record Linkage Practices. Researchers have been linking censuses together since the 1930s by manually matching individuals located on microfilmed enumeration forms. This laborious process yielded very small samples and was subject to substantial selection bias [4]. In December 2013, IPUMS released complete-count machine-readable census enumerations for nine census years from 1850 to 1940. These new historical data, covering almost 700 million individual records, are the fruit of collaboration between IPUMS and the world's two largest genealogical organizations—Ancestry.com and FamilySearch—to leverage genealogical data for scientific purposes [5]. Under agreements with those organizations, anonymized datasets are freely available to researchers, and restricted datasets including names and addresses are available under contracts that safeguard the proprietary interests of the genealogical organizations that donated the data.

As soon as the complete-count data with names became available in late 2013, economists and sociologists began developing methods for automated record linkage to construct longitudinal panels. At this writing, some 400 researchers are working on 148 research projects linking these datasets. This work has resulted in a flood of new research papers assessing the impact of early-life environmental and health conditions on later-life outcomes, as well as the

effects of education policy, multi-generational economic mobility, and the causes and consequences of migration [4].

The ferment of new research is exciting, but serious technical problems have arisen. Few investigators are well versed in advanced record linkage methods, and few have access to the software or high-performance computing needed for *reliable* large-scale linkage. Most studies use off-the-shelf statistical packages to do the matching, forcing compromises that reduce the reliability of the links. Moreover, most analyses are necessarily small scale, focusing on such groups as Norwegian immigrants, Southern migrants, persons born in a particular cohort, or persons known to have contracted childhood illnesses. Finally, the resulting datasets are sometimes proprietary, creating large barriers to new research in this area.

Record linkage is subject to two kinds of errors: false matches (or Type I errors) and missed matches (or Type II errors).[1] False match rates are the greatest concern, because these Type 1 errors introduce systematic upward biases in transition rates, such as migration rates, economic mobility, family transitions, or fluidity in racial identification. Suppose, for example, that an investigator seeks to measure migration. Falsely matched cases usually appear to be migrants, since two incorrectly linked individuals are unlikely to reside in precisely the same place.

Recent analyses have demonstrated that the most commonly used methods of automatic record linkage—which are usually based on phonetic classifications of names—have false match rates ranging from 20% to 70% [4, 7-8]. Such high error rates generally produce invalid estimates of transition rates. For example, Nix and Qian [9] randomly chose one match whenever their algorithm produced multiple tying matches. This approach gave them a very high match rate but an extremely high false-match rate, estimated by Bailey et al. at between 52% and 70% [7]. Nix

[1] In most of the computer science linkage literature, the level of Type I and Type II errors is measured according to the “precision” and “recall” metrics, which are expressed as success rates instead of error rates [6].

and Qian used their linked data to argue that 19% of black men "passed" for white at some time during their lives and that 10% then reverted to identifying as black. In view of the high potential for false matches in the Nix and Qian methodology, we view this result with skepticism.

Missed matches are also problematic, since they can introduce selection bias and reduce the representativeness of longitudinal panels. Most linkage studies measure missed matches poorly, using methods biased by mortality and outmigration. Analysts should measure missed matches relative to the population with the *potential to be matched*, which is the population present and alive in each source being linked. To mitigate the impact of selection bias resulting from missed matches, studies should then weight longitudinal panels to match key characteristics of the potentially linkable population, including ethnicity, economic status, education, lifetime migration, demographic characteristics, and family status. Few linkage studies execute these steps correctly.

Linking forward from IPUMS-MLP to modern data sources. Use of IPUMS-MLP will not be limited to historical analysis of change in the period from 1850 to 1940. Within restricted environments designed to protect privacy, investigators will be able to trace individuals from IPUMS-MLP forward to modern surveys, censuses, and administrative records. This capacity is critical, allowing a prospective view of population aging. IPUMS-MLP will provide access to the childhood environments and family backgrounds of people who are now old.[2] It will be easily linkable to modern survey data through an ongoing NIA-funded project to link five modern surveys of health and aging to the 1940 census and to modern administrative records, censuses, and surveys through the existing Census Longitudinal Infrastructure Project (CLIP). We are closely

[2] IPUMS-MLP uses only data sources that are fully in the public domain. Most of these records were made public by the National Archives and are now freely available through Internet search engines. As explained in the section on Human Subjects, the project poses no additional threat to privacy whatsoever. Under federal law, this includes federal censuses that are at least 72 year old; Social Security records for persons who have died; and military records from the two World Wars. We will also use historical vital records whose confidentiality varies from state to state.

coordinating our plans with these two ongoing data infrastructure projects to ensure that they will be able to seamlessly link IPUMS-MLP to modern sources. The potential of IPUMS-MLP to address current problems in aging and health is not confined to these two projects; rather they serve to illustrate the potential for transformative research offered by IPUMS-MLP.

Linking 1940 U.S. Census Data to Five Modern Surveys of Health and Aging. John Robert Warren is leading an NIA-funded effort (1R01 AG050300) to link records from the 1940 census to respondents of the Health and Retirement Study (HRS); the Panel Study of Income Dynamics (PSID); the Wisconsin Longitudinal Study (WLS); the National Social Life, Health, and Aging Project (NSHAP); and the National Health and Aging Trends Study (NHATS). These ongoing longitudinal studies are the cornerstones of America's data infrastructure for interdisciplinary research on aging and the life course, including topics such as physical and mental health, disability, and well-being; later-life work, economic well-being, and retirement; and end-of-life issues. A crucial weakness of these surveys is that they contain little information about social, economic, family, neighborhood, and environmental circumstances in childhood and young adulthood. The sparse early-life information now available was usually collected retrospectively, and the quality of these reports is largely unknown. This serious limitation of these data hinders researchers' ability to study the long-term impacts of childhood and young adult circumstances and to understand how later-life outcomes are the result of cumulative life-course processes. Linking these studies to the 1940 census vastly expands the analytic utility of the surveys for a variety of substantive problems. Because IPUMS-MLP will link the 1940 census backwards to 1930, 1920, and earlier censuses, the project will allow construction of a wide range of variables describing the family background of survey respondents through multiple generations, further enriching these crucial surveys.

Census Longitudinal Infrastructure Project (CLIP). CLIP is a major infrastructure project established by the Census Bureau’s Data Stewardship Executive Policy Committee in August

2014. Housed in the Census Bureau's Center for Administrative Records Research and Applications (CARRA) [10-12], CLIP is developing a general framework for longitudinal analysis of administrative and statistical records.

The 1940 census provides the baseline population for constructing millions of life histories in CLIP. The 1940 census is particularly valuable because it was the first to provide key indicators such as educational attainment and income and is the only census ever to include these inquiries for the entire population. In the CLIP infrastructure, designed data from censuses and surveys are linked to organic data from Social Security Administration, Medicare, and other administrative records. The first CLIP goal is therefore to assign Protected Identification Keys (PIKs) to each person in the IPUMS 1940 census file. The PIKs uniquely identify individuals across data sources for the purposes of improving data quality and program efficiency, and they are used within a secure Census Bureau computing environment to maintain confidentiality. To assign the PIKs, the Census Bureau has developed powerful tools for uniquely identifying individuals in census and administrative datasets, which serve as the underpinning for the CLIP linkage strategy. To date, CLIP has positively identified 72% of children age 0-9 who appear in the 1940 census [7]. CLIP data are allowing investigators to understand the origins of late life outcomes among people who reached age 65 between 1995 and 2005 and who are now aged 77 to 87; nine projects based on CLIP are already underway in Federal Statistical Research Data Centers. The Census Bureau obtained the machine-readable 1940 census from IPUMS, and therefore IPUMS-MLP will include exactly the same 1940 census records as CLIP. This will make it easy for the Census Bureau to join the two databases and create a powerful longitudinal resource spanning the period from 1850 to the present. Linked CLIP-MLP datasets will be available to researchers through the Federal Statistical Research Data Centers.

Research Opportunities. Linking IPUMS-MLP to modern surveys and administrative records will open extraordinary new opportunities for longitudinal research on population health and aging.[3] Thousands of innovative analyses will be feasible; consider the following use cases:

- *Impact of exposure to water-borne lead before age three on Late Onset Alzheimer's Disease (LOAD).* Prince [13] has suggested that lead might be an environmental source of a predisposition toward LOAD. This conjecture has been borne out in animal experiments but has not been assessed in human populations due to lack of information on lead exposure under age three, the period of greatest sensitivity identified in animal experiments. The IPUMS-MLP database can provide information about lead exposure in early childhood for millions of Medicare recipients through the well-known relationship between water pH and its plumbosolvency [14], and LOAD can be observed in the HRS or through CLIP linkages to Medicare records [15-16] and to the National Health Interview Survey.
- *Intergenerational transmission of health and well-being over multiple generations.* The relationship between outcomes for one generation (education, income, health) and those for earlier or later generations in the same family line has been examined in the U.S. almost exclusively in a two-generation (parent-child) context. This limitation is largely due to the paucity of data linking three or more generations [17-18]. With the 1940 Census as the keystone, IPUMS-MLP can be linked with the PSID or with IRS records through CLIP, permitting researchers to analyze both (1) the influence of up to six previous generations on an individual's outcomes, potentially revealing the extent to which two-generation research has understated the persistence in outcomes across generations, and (2) trends over more than 150 years in two-generation mobility.

[3] IPUMS-MLP will also enrich historical databases, especially the NIA-funded Early Indicators project (P01AG010120).

- *The receipt of Mothers' Pension benefits in childhood and later-life economic status, health, and well-being*. Aizer et al. [19] show a strong causal link between the receipt of Mothers' Pensions (the pre-1930 forerunner of AFDC/TANF) and a range of improved outcomes for children, such as healthy weight, higher income, and increased longevity. IPUMS-MLP and CLIP will together allow us to identify additional effects of early-life income support across the life course—on cognition, labor force participation, asset accumulation, and health at older ages—as well as the mechanisms through which these effects are produced.
- *The impact of early-life cognitive capacity on later-life health and economic outcomes*. Research into the role of cognitive measures on later-life earnings, health, and longevity in the U.S. has not been possible for large populations because no records connect cognitive testing early in life to census or administrative records on later life outcomes. There are machine-readable IQ scores for 500,000 U.S. male enlistees in World War II [14] that will be incorporated into IPUMS-MLP, allowing investigators to use CLIP to follow these individuals through census and health records from their childhood through late life and death.

Approach

When linking backwards from a more recent census to an older one, we can generally identify the potentially linkable population in the more recent census. The potentially linkable population between any pair of censuses is defined as the population of the terminal census year that was old enough to be present in the initial census year and that did not immigrate between the two census years. There are five main reasons for failing to locate potentially linkable persons in the earlier census: (1) Name changes: especially common when women marry but which may also occur because of name Anglicization or for other reasons; (2) Enumerator error in recording names or other characteristics; (3) Census underenumeration; (4) Transcription error; and (5) Multiple valid links: two or more persons exist with similar or identical linking characteristics. Because there are multiple opportunities for errors to be introduced, the linking algorithm must

accommodate approximate matches on a probabilistic basis. Planning and design of the linking algorithm must consider not only optimization of links but also computational efficiency; some techniques are extraordinarily computationally intensive and would be infeasible for a project of this scale if used on their own [6].

The sections that follow begin with description of our preliminary studies. We then explain the innovative record linkage strategies needed to produce a longitudinal resource of unprecedented scale. We present our new linkage technologies and algorithms. We then describe our approach to software development, dissemination, sustainability, metadata, project management, evaluation, and deliverables.

Preliminary Studies. The IPUMS-MLP research team has been engaged in developing automatic record linkage technology for the past two decades. Our most substantial previous linkage project completed is the IPUMS Linked Representative Samples (IPUMS-LRS) [20-21].

From 2002 until 2014, only one complete U.S. census enumeration—the 1880 census—was available to the research community. Over the course of two decades, volunteers of the Church of Jesus Christ of Latter-Day Saints (LDS) transcribed the entire population of the U.S. enumerated in the 1880 Census, a total of 50 million records. Working in collaboration with LDS, IPUMS converted this transcription into a source suitable for demographic research by correcting errors and coding millions of different alphabetic strings describing population characteristics into numeric categories [22]. The complete-count LDS database created an opportunity to link the 1880 census to historical IPUMS census samples covering 1% of the population for the period 1850 through 1930.

IPUMS-LRS designed procedures to minimize false links and maximize the representativeness of the linked cases, instead of maximizing the *number* of links. Most prior record linkage efforts, we argued, had focused too much attention on the percent of persons

missed by record linkage (Type II errors) and not enough on the percent of false links (Type I errors). Failing to identify links can lead to selection bias, but investigators can measure that bias and mitigate it by applying weights to make the linked cases representative with respect to observed characteristics. False matches, by contrast, can lead to systematic upward bias in migration rates, occupational mobility, and all kinds of family transitions. The IPUMS linked samples aimed to maximize both accuracy and representativeness. To minimize selection bias with respect to key transitions such as migration or widowhood or occupational change, we focused on a limited set of characteristics not expected to change over the life course: name, birth year, sex, and birthplace.

IPUMS-LRS linked datasets on a probabilistic basis using the Jaro-Winkler string comparison metric developed by the Census Bureau [23] and a machine learning algorithm known as a Support Vector Machine [24-25]. The machine learning software was "trained" with a set of hand-linked data developed by IPUMS staff. The probabilistic machine-learning software (1) compared every person of a given cohort, birthplace, and sex with every other person (in a census sample) that shared those characteristics and (2) predicted the probability of a true match based on similarity scores of several features, such as spelling of first name and last name, first and middle initials, phonetic name codes, name commonness, and age. By adopting conservative linking thresholds, IPUMS-LRS minimized false matches [20-21, 26]. Under the IPUMS-LRS linking procedure, whenever more than a single potential match was found, all potential matches were eliminated from consideration. To weed out false matches, IPUMS-LRS developed two models. The "loose" model was designed to maximize the number of potential links. The "tight" model was more selective, and established matches only where the fit was extremely close. Links were designated as true only if there was one and only one positive link in both models [20,27]. This approach sacrifices valid links to minimize false links and maximize representativeness. The loose model excludes cases with an observed possibility of choosing the wrong match, and the

tight model excludes cases with any significant discrepancies in name or age. The samples were weighted to approximate the characteristics of the potentially linkable population with respect to family relationship, birthplace, age, size of place, and occupation. The conservative design of IPUMS-LRS yielded false match rates well below 5%, which is low relative to prior intercensal linkage procedure [4, 7, 20]. Lifetime interstate migration (from birth to the census) and family transition rates are consistent with estimates based on other methods, providing strong evidence that IPUMS-LRS yields unbiased and reliable estimates of life-course transitions [20, 26]. IPUMS-LRS was the first project to implement a machine learning approach to historical census record linkage and provided invaluable lessons for the larger project we are now implementing [4].

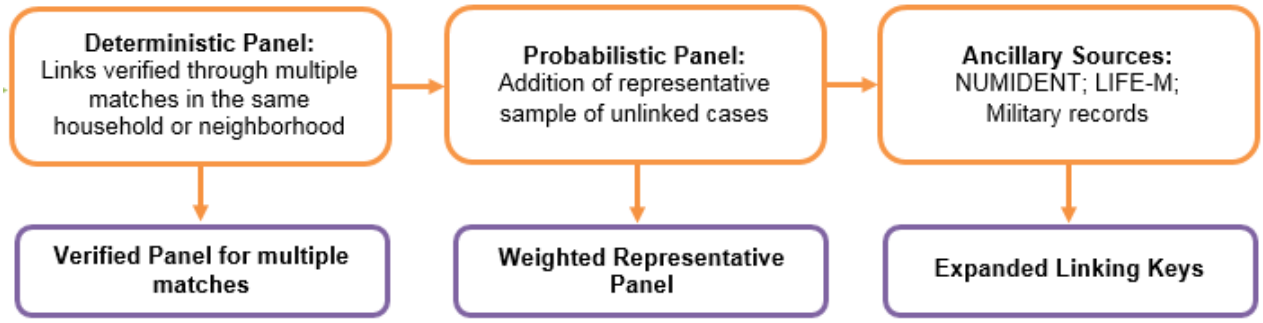
IPUMS-LRS has limitations that limit its usefulness for longitudinal analysis of health and aging. Most important, only two observations are available for each linked individual. Moreover, because the strategy relied on census samples, the linked datasets are small. The new IPUMS-MLP will include a thousand times as many links across nine complete-count censuses, Social Security, military, and vital records.

IPUMS-MLP also builds on the work of two other research projects that are currently engaged in historical record linkage research. A project under the leadership of J. David Hacker (R01 HD082120) is developing models of demographic and health changes following military conflict. As part of this analysis, Hacker is working on linkages of the 1850 through 1880 complete count censuses. The software developed for this linkage work will served as the foundation for the IPUMS-MLP linkage software. As described below, have enhanced the Hacker software by improving computational efficiency and implementing new linking strategies. The second project is the Longitudinal Intergenerational Family Electronic Micro-Database (LIFE-M) project underway at the University of Michigan under the direction of Martha Bailey [28]. The goal of LIFE-M is to link records of births, marriages, and deaths for people born between 1880 and 1930 to

construct life histories of demographic events. As explained below, LIFE-M will help IPUMS-MLP with the difficult problem of linking women who marry and change names.

Linking Strategy. We plan a multi-stage linking process, with three major components: (1) A deterministic linked panel based on cases verified by multiple linked persons within a household and/or multiple linked persons within an immediate neighborhood. Our preliminary analysis demonstrates that this deterministic panel will match approximately 50% of cases with virtually zero false matches. (2) A probabilistically linked representative sample of the remaining unlinked population, using machine-learning technology that capitalizes on the deterministic links for training and evaluation. (3) An augmented panel that adds information from administrative records to enable linking of women who marry in the interval between censuses and enrichment of the database with additional variables. Figure 1 provides a schematic of the three processing stages.

Figure 1. Multigenerational Record Linkage Workflow



1. Deterministic Panel. The deterministic panel will consist of cases in which there are multiple matches within a particular household or among immediate neighbors. Matching multiple people within the same unit effectively eliminates false matches. One of the greatest challenges of intercensal record linkage is that there are often multiple potential matches for a given individual. For example, the 1880 census lists 44 white men named John Smith who were born in New York State in 1848. Just one of these John Smiths was married to a Mary Smith and had a

son who was also named John Smith. If we add the fact that Mary was born in Wisconsin in 1854 and John Jr. was born in 1878, the odds of a coincidental match are infinitesimal. This example is just about the worst-case scenario, since John and Mary Smith were the two most common names, and New York was the most common state of birth. The combination of name, age, sex, race, and birthplace for a group of at least three individuals yields trillions of permutations, which would be virtually impossible to match by chance.

We are finalizing rules for these matches based on blocking strategies (dividing the population into subgroups based on broad similarity), Jaro-Winkler string-comparison metrics for names, and rules governing the similarity of birth years and birthplaces. This approach is computationally demanding since it involves trillions of comparisons; as described below, however, we have now overcome the problem of scale through the application of innovative technology, and we expect to process the links on the IPUMS high-performance computing cluster. We will begin by matching each pair of decades, working backward from 1940 (i.e., 1940-1930, 1930-1920, and so on). We will then make matches over 20-year intervals to capture cases that were missing or poorly enumerated in a particular census year (e.g., 1940-1920, 1930-1910, and so on). We will create an identity key, a unique identifier for each individual who appears in more than one census year, which will provide a linking key to match persons over multiple census years.

The deterministic procedure can link approximately half of individuals without a significant number of false matches. The resulting panel, however, is unrepresentative, since it systematically excludes persons who migrate without family members, a group that represents a significant and theoretically important segment of the population. The verified deterministic panel will be suitable for many purposes, such as analysis of fertility limitation and child mortality. It is less suitable for other applications, since it will tend to underrepresent migration, occupational mobility, and family dissolutions.

2. Probabilistic Representative Panel. The second iteration of the linked panel will address the problem of representativeness. We will use probabilistic machine learning procedures to link a representative subset of the cases that we were unable to link in the deterministic pass. Our approach in this phase is conceptually similar to IPUMS-LRS, but we will rely on newly developed technology to maximize efficiency and accuracy. Following the procedures of IPUMS-LRS, linkages in this stage will be based exclusively on characteristics that are not expected to change over the life-course: names (except for women who marry), birth year, sex, race, birthplace, and birthplaces of parents.[4] For some census years, we can also use information on year of immigration, mother tongue, or age at first marriage. We will not use any characteristics that are expected to change over the life course, such as place of residence or the characteristics of co-resident family members.

The probabilistic panel will rely on machine-learning technology (described below), so we need accurate training data so that the software can learn the combinations of characteristics associated with true matches. The deterministic panel will provide a massive set of highly-reliable training data. Because we will have complete enumerations rather than samples at both ends of each linked pair of censuses, we will be able to identify and exclude multiple matches at either end of any linked pair of censuses and thus further reduce the low false-positive rate achieved by IPUMS-LRS. We will filter the results to eliminate matches with contradictory information, such as a completely inconsistent set of kin in the two censuses. Despite this aggressive elimination of false matches, we anticipate that our new technology (described below) will successfully match approximately 20% of the previously unlinked cases.

As noted, the potentially linkable population between any pair of censuses is the

[4] Respondents may report any of these characteristics differently in different censuses. For example, they may change their names for reasons other than marriage, and some change their racial or ethnic identities [29]. The matching algorithms described below can accommodate imperfect matches, but fluid responses will reduce linkage rates.

population of the terminal census year that was old enough to be present in the initial census year and that did not immigrate between the two census years. In most intervals, we can directly identify the immigrants using information in the census; for the rest, we can develop estimates of the potentially linkable population using immigration statistics. The Representative Panel will include all the linked individuals in the Deterministic Panel, and thus non-migrants and persons with stable families will be overrepresented. To overcome this problem, we will weight the linked representative panel by iterative proportional fitting (raking) to match the characteristics of the potentially linkable population with respect to age, sex, race, family relationship, presence of kin, state/country of birth, lifetime migration, size of place, and occupation.

3. Linked Administrative Records. The census-only linked panel has limitations. Most important, it loses women who marry and change their names during the interval between censuses. By linking to additional sources for our third panel, we can achieve more accurate and comprehensive links between censuses. In addition, other sources can provide additional variables as well as extra detail on events occurring between censuses.

Our core administrative source will be a new public version of the Social Security Applications and Claims Index (Numident) available through the National Archives [30]. The public version of the Numident includes persons with a Social Security number who either had died or who would have reached age 110 by 2007. The Numident records include each Social Security applicant's full name, maiden name, exact dates of birth and death, place of birth, place of death, citizenship, sex, father's name, mother's maiden name, and race/ethnic description. We will use the Numident to follow women, even those who marry in the interval between two censuses, from their families of origin to the families in which they reside as adults. Thus, a 40-year-old woman observed in the 1940 census can be located in her parents' household in the 1910 census. Because her mother's maiden name is also given in the Numident and her mother's birthplace and birth year are known from the 1910 census, the mother can in turn be located in the 1880

census. The Numident will dramatically improve linkage rates for married women, and because linking can be done with a public version of the Numident, links generated in this way can be publicly disseminated without restrictions.

A second key administrative source will be the Longitudinal Intergenerational Family Electronic Micro-Database (LIFE-M) currently under development at the University of Michigan [28]. LIFE-M is reconstituting families by linking vital records. Beginning with birth certificates from 1881 to 1930, LIFE-M is matching the birth records to marriage records and death records. By constructing family histories across multiple generations, LIFE-M will allow researchers to study processes of demographic change in unprecedented detail. We can link LIFE-M to IPUMS-MLP using information on place of residence at the census nearest to a vital event, as well as name, spouse's name, birth year, birthplace, and sex, all of which are available in both sources. Once we have established links between LIFE-M and IPUMS-MLP, we can use the LIFE-M data to refine and augment the IPUMS-MLP links, particularly for women who change their names at marriage. In addition, the availability of exact birthdates in LIFE-M will also aid in disambiguating links to the 1900 census, which provides month and year of birth, and to the Numident, which provides exact birthdate. LIFE-M will not only help to improve IPUMS-MLP; in addition, links from IPUMS-MLP will flow back to LIFE-M. The chief limitation of LIFE-M is that the availability of vital records limits the states and periods that it covers, but we anticipate that the scope of the database will expand as additional vital records become available in machine-readable form.

We will also link IPUMS-MLP to draft and enlistment records from World War II and draft records from World War I [31-32]. These records can be linked using information on place of draft registration as well as name and birth year, and the military sources will provide additional variables, including exact date of birth and exact place of birth. (The census records only state of birth for the U.S.-born or country of birth for the foreign-born).

Linking Process. Our linking process will consist of the following elements: name cleaning, blocking, string comparison, machine learning, high-performance processing, and assignment of identity keys.

Name cleaning. Record linkage begins with software for parsing and standardizing names. While names are the most important piece of information available for record linkage, they are also the most problematic. Errors in naming can arise from respondent error (as when, for example, a farm wife responding to an enumerator misstates the name of a farm hand), enumerator error, or transcription error. Moreover, names often change over time; they sometimes lack standard spelling; and in some cases, people were enumerated using a nickname or middle name in one census and a formal first name in another.

To minimize error from these sources, we are implementing a comprehensive program of name cleaning, accounting for common typographical transpositions and handwriting recognition errors. We standardize given names to account for diminutives and abbreviations (e.g., "Willie" and "Wm." are transformed into "William"). This work draws on a rich body of research on name cleaning [33-37]. We also employ phonetic name coding, a standard tool for record linkage since the 1930s. The most commonly used systems are Soundex, NYSIIS, and Phonex. We rely mainly on the Double Metaphone system, which returns two encoded strings corresponding to variant pronunciations [38-39].

Blocking. Every pair of records drawn from two files is either a match referring to a single individual or is a non-match describing two different persons. Optimal matching requires that every individual be compared with every possible match [40]. It is not computationally feasible, however, to evaluate every potential match; implementation of such a linking algorithm for the full 1850-1940 datasets would involve about 2.5×10^{17} comparisons. To reduce the computational requirements, we introduce "blocking factors" that define a subset of the population and limit

comparisons to persons who share the same blocking factors. For IPUMS-LRS, the blocking factors were state of birth, sex, and a seven-year window of birth year.

Our most important means of improving record linkage efficiency is to greatly reduce the size of the blocks. We will do this mainly by classifying first and last names into blocks. The danger of small blocks is that they may miss true matches. Thus, for example, a one-character transcription error could eliminate a valid match from consideration in a blocking system based solely on first initials. Accordingly, we are designing multiple overlapping small blocks that in combination capture all plausible links. We have shown that blocking by surname bigrams—successive pairs of letters—we can capture nearly 100% of potential matches with a 75% reduction in the number of comparisons needed. We are continuing our experimentation and expect to improve on these results [41].

String Comparison. As we did for IPUMS-LRS, we plan to use the Jaro string comparator as modified by Winkler for name comparison [28]. This algorithm computes a similarity measure between 0.0 and 1.0 based on the number of common characters in two strings, the lengths of both strings, and the number of transpositions, accounting for the increased probability of typographical errors towards the end of words. The other linking variables—birthplace, parental birthplaces, age, sex, and race—pose few string comparison problems because those variables are already classified and numerically coded according to the IPUMS coding system. Thus, for example, we will not have to cope with the innumerable spelling variations of Massachusetts. We will, however, develop an algorithm for age misreporting that can account for digit preferences: inconsistencies in age between two census years should be partly discounted if age is rounded to a five or zero in one or both census years.

Machine Learning. We will implement machine learning algorithms to optimize the quality of links. IPUMS-LRS used an algorithm known as a Support Vector Machine (SVM) to classify each possible match [41-44] using an open-source library of tools developed by Chang and Lin

[24]. For IPUMS-MLP, we will test newer machine learning strategies to identify the optimal solution. Our preliminary testing shows that a random forest has the potential to reduce false positives by two percentage points [45]. To optimize linkage, we will test cutting-edge strategies of random forests (an ensemble classification method based on many decision trees) as well as unsupervised collective graph matching [46]. Finally, we will assess recently proposed active learning techniques for record linkage [47], which have out-performed fully-supervised SVM and Decision Trees in recent matching tests.

To estimate the matching parameters, we need training data, namely, cases where the true links are known. We plan to use training data developed from genealogical sources. To estimate error rates in the record linkage, we divide the training data in two, using part to estimate the parameters for machine learning and the rest to test the linking algorithm to estimate both the rate of false positive matches (Type I errors) and omitted true matches (Type II errors).

High Performance Record Linkage Technology**.** The project will develop and implement technical solutions that can accommodate the massive scale of the database. Our largest previous linkage project, IPUMS-LRS, was tiny compared with IPUMS-MLP. IPUMS-LRS was nevertheless highly computationally-intensive, requiring the use of 900,000 core-hours of supercomputing time provided by the Minnesota Supercomputing Institute using a Silicon Graphics Altix XE 1300 Linux cluster with 2048 compute nodes and 4TB of main memory. IPUMS-MLP will include approximately 2,000 times as many links as IPUMS-LRS. All things being equal, the number of computing operations required for record linkage is proportional to the square of the number of links being processed. If we used the same procedures as for IPUMS-MLP as we did in IPUMS-LRS, we would need about four million times as much computing power. That scale of computing simply does not exist. Accordingly, we have taken steps to improve the *efficiency* of our linking strategies.

We have developed new highly-efficient linking software, known as Hlink [47]. HLink was designed to overcome limitations of existing linkage software. Existing record linkage tools are available as libraries for common data manipulation languages such as R, Python, and Java. These libraries present an array of features such as preprocessing techniques, indexing techniques, supervised and unsupervised learning techniques, and rudimentary graphical user interfaces. The existing software creates features from datasets that exist of individual records, and a large amount of custom data manipulation is required beforehand to use contextual household information [49]. Moreover, existing software makes it difficult to use multiple nodes and little support for parallel processing [50].

To overcome these limitations, Hlink provides a single end-to-end linking solution, replacing a multi-stage process used for IPUMS-LRS that involved multiple programs (e.g., FEBRL, LIBSVM, C, statistical packages) and data formats (e.g., flat ASCII files, MySQL, binary files). The system leverages Apache Spark, a Hadoop-based technology. Spark enables parallel processing in all stages of record linkage. We ingest the data into Apache Parquet, a columnar storage database format that improves read access for sequential files compared with conventional relational databases. We use Apache Spark's support for the Parquet storage format tied to a record shredding and assembly algorithm to optimize parallel processing [51-53]. There are small pieces of functionality written in Scala for custom transformations that were not available in the standard Spark library. The database was created with Apache Hive, which is built into Spark, and all files are stored using the parquet file format. Using these technologies, we have already improved query execution speed by two orders of magnitude. These innovations—in combination with the efficient new blocking strategies described above—give us the computing power we need to construct the world's largest longitudinal population panel.

Identity Keys**.** A unique identity key will consistently identify individuals in every sample in which they are found. At each stage of record linkage, we will begin with the 1940 census; linking

backwards from more recent censuses to earlier ones avoids linkage failure due to mortality or emigration. We will assign an identity key to each individual age 10 or older in the 1940 census. We will link 1940 backwards to each of the prior census years, assigning Identity keys to individuals in the prior censuses whenever we successfully establish a link.

We will then turn to the 1930 census, successively linking backwards from 1930 to each previous census year. When we establish a new link, we will assign the 1940 Identity key whenever it is available. If no Identity key exists (because the individual in 1930 was not linked to 1940), we will generate a new key as needed. In this fashion, we will proceed backwards from census to census until we have established links between every possible combination of census years. At the conclusion of this process, there are bound to be some inconsistencies. We will reconcile all inconsistencies using rule-based and probabilistic strategies, ensuring that all keys identify only a single individual across all censuses.

For each individual, we will also construct variables identifying the linking keys for spouses, mothers, fathers, up to four grandparents, and up to eight great-grandparents. These interrelationship keys will be available in all census years; so, for example, a future spouse will be identifiable when the individual is still a child. Siblings will be identified because they share a parental linking key, cousins because they share a grandparent key, and so on. Accordingly, the family interrelationship variables will not only be valuable for assessing family influences across multiple generations but will also allow study of lateral kin relationships beyond the household.

Dissemination. Data sharing is central to the project: effective dissemination is essential if the data are to be widely used. Each data release will appear in two formats: (1) a public version without names or character strings available at ipums.org and (2) a restricted version with names and addresses available by license. We also anticipate that the Census Bureau will make the data available to researchers through the Federal Statistical Research Data Centers (FSRDCs) once it is linked with recent administrative records and surveys.

Under our agreement with Ancestry.com, the original string data (including the names) are limited to access through restricted licenses. Accordingly, the public use IPUMS-MLP will not include names or character strings of any kind; all variables will be numerically coded. The restricted version, with full names and addresses, will be made available to researchers who sign a restricted data license and agree to keep the data highly secure.

We will distribute the public data and documentation through a web-based data access system that will construct customized longitudinal files designed according to user specifications. Because of the large scale of the data, we must provide efficient subsetting and data manipulation capabilities. We plan to build a web-based dissemination system that will produce linked extracts in a variety of formats incorporating any combination of census years. The IPUMS data access system pioneered web-based dissemination of large-scale datasets, and the IPUMS dissemination tools continue to innovate at the cutting edge of information technology. We will leverage that software for IPUMS-MLP. The system for disseminating the linked data will offer advanced capabilities for navigating documentation and defining datasets that capitalize on the longitudinal structure of the data.

We will disseminate customized datasets as ASCII text files and in the proprietary formats of the major statistical packages (Stata, SAS, SPSS, and R). We will provide online data analysis tools for those who do not wish to download the data. With each dataset, we will provide a customized codebook in Data Documentation Initiative (DDI) structured XML format with a stylesheet allowing users to view the codebook in a web browser. We will also disseminate more comprehensive documentation (such as instructions for enumeration of each variable) through our data access system. We will release all software used for the project under an open-source license, both for purposes of documentation and to enable researchers to apply our linking methods to other datasets.

References Cited

1. Groves R. 2011. Three Eras of Survey Research. *Public Opinion Quarterly.* 75: 861-871.
2. Warren JR, Hauser R. 1997. Social Stratification across Three Generations: New Evidence from the Wisconsin Longitudinal Study. *American Soc. Rev.* 62: 561-572.
3. Mare RD. 2011. A Multigenerational View of Inequality. *Demography* 48: 1-23.
4. Ruggles, S, Fitch C, Roberts E. 2018. Historical Census Record Linkage. *Annual Review of Sociology* 44:19-37.
5. Ruggles S. 2014. Big Microdata for Population Research. *Demography,* 51: 287-297.
6. Christen P. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. New York: Springer
7. Bailey MJ, Cole C, Henderson M, Massey C. 2017. How well do automated linking methods perform in historical samples? Evidence from new ground truth. Work. Pap., Dept. Econ., Univ. of Michigan
8. Massey CG. 2017. Playing with Matches: An Assessment of Accuracy in Linked Historical Data. *Historical Methods*. 50(3):1–15
9. Nix E, Qian N. 2015. The Fluidity of Race: "Passing" in the United States, 1880-1940. Work. Pap. 20828, National Bureau of Economic Research
10. Alexander JT, Gardner T, Massey CG, O'Hara A. 2015. Creating a Longitudinal Data Infrastructure at the Census Bureau. CARRA Working Paper 2015-1, U.S. Census Bureau.
11. Massey CG. 2014. Creating Linked Historical Data: An Assessment of the Census Bureau's Ability to Assign Protected Identification Keys to the 1960 Census. CARRA WP 2014-12, U.S. Census Bureau.
12. Massey CG, O'Hara A. Person Matching in Historical Files using the Census Bureau's Person Validation System. CARRA WP 2014-11, U.S. Census Bureau.
13. Prince M. 1998. Is Chronic Low-Level Lead Exposure in Early Life an Etiologic Factor in Alzheimer's Disease? *Epidemiology* 9: 618-621.
14. Ferrie JP, Rolf K, Troesken W. 2012. Cognitive Disparities, Lead Plumbing, and Water Chemistry: Prior Exposure to Water-Borne Lead and Intelligence Test Scores among World War Two U.S. Army Enlistees. *Econ. Human Biology.* 10: 98–111.
15. Taylor, Donald H., G.G. Fillenbaum, M.E. Ezell. 2002. The Accuracy of Medicare Claims Data in Identifying Alzheimer's Disease. *Journal of Clinical Epidemiology.* 55(9): 929-937.
16. Taylor, Donald H., Truls Østbye, Kenneth M. Langa, David Weir, and Brenda Plassman. 2009. The Accuracy of Medicare Claims as an Epidemiological Tool: The Case of Dementia Revisited. *Journal of Alzheimer's Disease*, vol. 17, no. 4, pp. 807-815.
17. Pfeffer FT. 2014. Multigenerational Approached to Social Mobility: A Multifaceted Research Agenda. *Research in Social Stratification and Mobility* 35: 1-12.
18. Ferrie, Joseph, Catherine Massey, Jonathan Rothbaum. 2016. Do Grandparents and Great-Grandparents Matter? Multigenerational Mobility in the US, 1910-2013. NBER Working Paper No. 22635.

19. Aizer A, Eli S, Ferrie JP, and Lleras-Muney A. 2016. The Long-Run Impact of Cash Transfers to Poor Families. *American Econ. Rev.* 106: 935-971.
20. Goeken R, Huynh L, Lynch TA, Vick R. 2011. New Methods of Census Record Linking. Historical Methods. 44(1):7–14
21. Ruggles S. 2002. Linking Historical Censuses: a New Approach. *History and Computing.* 1+2 (publ. 2006):213–24
22. Goeken R, Nguyen C, Ruggles S, Sargent W. 2003. The 1880 U.S. Population Database. *Historical Methods* 36: 27-34.
23. Porter, E.H., &Winkler, W.E. 1997. Approximate String Comparison and its Effect on an Advanced Record Linkage System. Census Bureau Research Report RR97/02. Washington
24. Chang C, Lin C. 2007. LIBSVM: A Library for Support Vector Machines. Department of Computer Science, National Taiwan University. http://www.csie.ntu.edu.tw/~cjlin.
25. Christen P. 2008. "Febrl—A Freely Available Record Linkage System with a Graphical User Interface." Warren, JR, Yu P, Yearwood J, editors. In *Conferences in Research and Practice in Information Technology.* http://crpit.com/confpapers/CRPITV80Christen.pdf.
26. Ruggles S. 2011. Intergenerational Coresidence and Family Transitions in the United States, 1850-1880. *Journal of Marriage and the Family,* 73: 138-148.
27. Johnson DS, Massey C, O'Hara A. 2015. The Opportunities and Challenges of Using Administrative Data Linkages to Evaluate Mobility. *Annals of the American Academy of Political and Social Science.* 657(1):247–64
28. Bailey MJOURNAL 2017. The Longitudinal, Intergenerational Family Electronic Micro-Database Project. Ann Arbor, MI: Univ. Michigan.
29. Liebler, C.A., Porter, S.R., Fernandez, L.E., Noon, JOURNALM. and Ennis, S.R., 2017. America's Churning Races: Race and Ethnicity Response Changes between Census 2000 and the 2010 Census. *Demography*, 54(1), pp.259-284.
30. National Archives and Records Administration. 1936-2007. Numerical Identification Files (NUMIDENT). Record Group 47, Records of the Social Security Administration. Electronic and Special Media Records Services Division, College Park, MD.
31. National Archives and Records Administration. 2002. U.S. Army Serial Number Electronic File. (1938-46). Record Group 64, Records of the National Archives and Records Administration. Electronic and Special Media Records Services Division, College Park, MD.
32. National Archives and Records Administration. 1917-1918. World War I Selective Service System Draft Registration Cards. Publication Number: M-1509. Electronic version available through Ancestry.com.
33. Christen P, Churches T, Zhu JOURNAL 2002. Probabilistic Name and Address Cleaning and Standardisation. *Proceedings of the Australasian Data Mining Workshop*, December, Canberra, Australia.
34. Winkler WE. 1990. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *American Statistical Association 1990 Proceedings of the Section of Survey Research Methods,* 354-359.
http://www.amstat.org/sections/srms/Proceedings/papers/1990_056.pdf
35. Nygaard L. 1992. Name Standardization in Record Linkage: an Improved Algorithmic Strategy. *History and Computing,* 4: 63-74.

36. Maletic JI, Marcus A. 2000. “Data Cleansing: Beyond Integrity Analysis.” Pp. 200-209 in *Proceedings of the Conference on Information Quality*. Boston: Massachusetts Institute of Technology.

37. Vick R, Huynh, L. 2011. The Effects of Standardizing Names for Record Linkage: Evidence from the United States and Norway. *Historical Methods,* 44: 15-24.

38. Philips L. 2000. The Double-Metaphone Search Algorithm. *C/C++ User's Journa*l 18. http://drdobbs.com/cpp/184401251

39. Lait AJ, Randell B. 1993. An Assessment of Name Matching Algorithms. Department Technical Report Series No. 550, Department of Computing Science, University of Newcastle upon Tyne, UK.

40. Fellegi IP, Sunter AB. 1969. A Theory for Record Linkage. *Journal of the American Statistical Association,* 64: 1183-1210.

41. Christen P. 2012. A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication. *IEEE Transactions on Knowledge and Data Engineering,* 24(9):1537-1555.

42. Christiani N, Shawe-Taylor J. 2000. *An Introduction to Support Vector Machines.* Cambridge, England: Cambridge University Press.

43. Abe S. 2005. *Support Vector Machines for Pattern Classification.* London: Springer-Verlag.

44. Vapnik VN. 1998. *Statistical Learning Theory.* New York: Wiley Interscience.

45. Christen, P., 2012. A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication. *IEEE Transactions on Knowledge and Data Engineering*, 24(9), pp.1537-1555.

46. Fu Z, Boot M, Christen P, Zhou J. 2014. Automatic Record Linkage of Individuals and Households in Historical Census Data.” *International Journal of Humanities and Arts Computing,* 8: 204-225.

47. Christen P, Vatsalan D, Wang Q. 2015. Efficient Entity Resolution with Adaptive and Interactive Training Data Selection. *IEEE International Conference on Data Mining (ICDM '15)*, Atlantic City, November 2015, 727-732. DOI 10.1109/ICDM.2015.63

48. Wellington, J. 2018. HLink: An Investigation into Historical Record Linkage. M.S. Thesis, University of Minnesota.

49. De Bruin, J. N.D. Data Matching Software. https://github.com/J535D165/data-matching-software

50. Enamorado, Ted. Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records. 2017, https://imai.fas.harvard.edu/research/linkage.html.

51. Apache Parquet. 2016. Retrieved 15 January 2016 from parquet.apache.org.

52. Armbrust M, Xin RS, Lian C, Huai Y, Liu D, Bradley JK, Meng X, Kaftan T, Franklin MJ, Ghodsi A, Zaharia M. 2015. “Spark Sql: Relational Data Processing in Spark.” In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data,* 1383–1394.

53. Melnik S, Gubarev A, Long JJ, Romer G, Shivakumar S, Tolton M, Vassilakis T. 2011. Dremel: Interactive Analysis of Web-Scale Datasets.” *Communications of the ACM,* 54: 1