IPUMS Redesign

STEVEN RUGGLES
MATTHEW SOBEK
MIRIAM L. KING
CAROLYN LIEBLER
CATHERINE A. FITCH
Minnesota Population Center
University of Minnesota

Abstract. This new project will create two large parallel series of historical U.S. census microdata. The first is a redesigned Integrated Public Use Microdata Series (IPUMS) incorporating Census 2000 and the American Community Survey. The IPUMS is a compatible series of large census microdata samples spanning the period from 1850 to 2000. The second is a restricted-use microdata archive containing 1.4 billion records from the censuses of 1940 to 2000. The new restricted-use archive is the product of a Census Bureau initiative to harmonize all of the bureau's decennial microdata and make them accessible through the Census Bureau Research Data Centers. The project will collaborate with the Census Bureau to create coding schemes to be used in both series that incorporate all detail in the restricted files. The two series will be developed simultaneously using the same software, methodology, and documentation.

Keywords: census, Census Bureau, IPUMS, microdata

Overview

IPUMS Redesign is an ambitious new project to overhaul the U.S. component of the Integrated Public Use Microdata Series (IPUMS). The project began in October 2002 with funding from the National Institutes of Health (HD-43392). The IPUMS is a compatible series of large U.S. census microdata samples spanning the period from 1850 to 1990. Originally designed in the early 1990s, the IPUMS must be extensively modified to accommodate new data sources and to keep up with changing technology and metadata standards. The project is still in its early stages, and our plans are subject to change. We are publishing this description in the hope that data users will respond with comments and suggestions that will help us to better meet the needs of IPUMS users. A complete copy of the proposal is available from the IPUMS Web site (http:// www.ipums.org).

The IPUMS Redesign project has three components. First, we will create a new version of the IPUMS that includes data from Census 2000 and the American Community Survey (ACS). The addition of these new data files will require modification of several IPUMS coding schemes to maximize cross-census compatibility. Second, we will assist the U.S. Census Bureau in the creation of a vast new IPUMS-compatible restricted-use microdata archive, including 1.4 billion records from the censuses of 1940 through 2000. This data archive—which includes complete long-form census microdata for each census since 1960 and short-form data for 1970 onward—will be available in IPUMS format through the Census Bureau Research Data Centers. Finally, we will overhaul the infrastructure underlying the database. This will include converting the IPUMS documentation to new metadata standards, improving the functionality of the data access system, creating a version-control protocol, and revamping the IPUMS software.

Expanded and Improved IPUMS Public Files

As described in the introduction to this theme issue, the IPUMS has become one of the most widely used data sources in American social science. The substantial use of the IPUMS is remarkable considering that the most recent census data included are now 12 years old. The incorporation of Census 2000 into the series will allow researchers to place issues of current policy concern in broader chronological perspective. The IPUMS will remain current as we add ACS—a large annual sample that replicates the census long form—to the data series beginning in 2004.

Table 1 summarizes the source files for the public use components of the data series. This series will span the period from 1850 through 2000 with census data and will be continued with annual ACS data from 2003 onward. According to current Census Bureau plans, the public use microdata from the ACS will share the same specifications as Census 2000 public microdata, so the design issues posed by Census 2000 will also apply to the ACS.

Because of heightened concern about respondent confidentiality, the Census 2000 microdata files will incorporate

TABLE 1. Principal Data Files in the Public Microdata Series

Census year	Description	Sample density	No. of records (in thousands)		IPUMS variables	IPUMS file size (megabytes)
			Household	Person		
Public files						
1850	General sample	.01	37	198	92	79
1860	General sample	.01	66	354	94	141
1870	General sample	.01	80	428	94	170
1880	General sample	.01	107	503	123	204
1900	General sample	.01	208	870	115	361
1910	General sample	.01	311	1,271	125	682
1920	General sample	.01	257	1,037	122	433
1930	General sample	.01	334	1,232	130	508
1940	General sample	.01	391	1,351	174	584
1950	General sample	.01	461	1,922	170	798
1960	General sample	.01	579	1,780	141	790
1970	State samples	.02	1,488	4,060	206	1,858
1970	County group samples	.02	1,488	4,060	210	1,858
1970	Neighborhood samples	.02	1,488	4,060	203	1,858
1980	Metro sample	.01	942	2,267	276	1,075
1980	Urban/rural sample	.01	942	2,267	266	1,075
1980	State sample	.05	4,711	11,337	276	5,376
1990	Metro sample	.01	1,106	2,500	252	1,208
1990	State sample	.05	5,528	12,500	252	6,039
2000	National sample	.01	1,132	2,814	252	1,353
2000	5% sample	.05	5,660	14,070	252	6,764
2003	ACS	.01	1,143	2,842	252	1,335
2004	ACS	.01	1,154	2,870	252	1,348
2005	ACS	.01	1,166	2,899	252	1,362
Total			30,779	79,492		37,259

Note. ACS = American Community Survey.

less detail than previous Census Bureau products. The Census Bureau will produce two public use microdata samples for Census 2000. The first of these—a national 1 percent sample released in spring 2003—includes the full subject detail that was available in the 1990 microdata sample but will offer sharply reduced geographic precision. The second product—a 5 percent sample scheduled for release in fall 2003—will have geographic detail similar to earlier censuses but will have reduced detail in many subject areas, including birthplace, ancestry, Hispanic origin, language, occupation, and migration. According to current Census Bureau plans, we can expect similar treatment of subject areas in the ACS. This reduction in detail renders the current IPUMS coding of those variables obsolete.

Along with changes in the public files resulting from confidentiality concerns, Census 2000 incorporates restructured variables describing race, occupation, and industry, as mandated by new guidelines from the Office of Management and Budget (OMB). Substantial revisions of several IPUMS coding schemes will be required to allow comparison of these subject areas across time.

Coding Design Principles

When the IPUMS was designed in 1991, we were at pains to minimize file size. Data storage costs were high, and because we had no automated capability to extract subsets of the data, we assumed users would obtain entire data files. During the past decade, the cost of random-access data storage has declined almost a thousandfold. Moreover, the emergence of the Internet, together with the Web-based data access systems pioneered by the IPUMS project, have changed the way users interact with data. These changes offer the potential for significant improvements in data structure and variable design.

To maximize temporal compatibility of variables with no loss of detail, the IPUMS employs composite coding systems for most complex variables. The first digits of the composite code provide information available across all samples. One or two additional digits provide added detail for a particular census year or group of years. For example, there is a two-digit general relationship code that provides the lowest common denominator that can be identified in all census years and a four-digit detailed relationship code that gives additional information available in a subset of years.

This approach does not always maximize cross-census compatibility. Because of the decline in data storage costs, the redesigned IPUMS can include specialized versions of variables, or of detail codes, that are optimized for particular combinations of census years. When creating a data extract, users will have the option of specifying that they want the maximum level of detail available for the particular combination of census years they are using. Users who prefer standard IPUMS codes, however, will still be able to obtain them.

Backward-Compatible Race Variables

As with prior censuses, in 1990 the census asked respondents to "Fill ONE circle for the race that the person considers himself/herself to be." Census 2000, however, instructed respondents to "mark *one or more* races to indicate what this person considers himself/herself to be" (emphasis added). For whites, the effect of the change is comparatively small; just 2.5 percent of those who checked the box for white also indicated another race. For other races, however, the multiple race option leads to major problems of historical compatibility: 4.8 percent of blacks, 39.9 percent of American Indians, 13.9 percent of Asians, and 54.4 percent of Native Hawaiians or Pacific Islanders indicated more than one race.

The shift from a single-race to a multiple-race census inquiry is a fundamental conceptual change, and we cannot construct a perfectly backward-compatible variable. We can, however, provide researchers with a range of constructed race variables that maximize historical compatibility for specific research applications.

The OMB (Office of Management and Budget 2001) identified several alternate strategies for bridging the change in the definition of race. The goal is to identify a statistical model predicting as closely as possible the responses that would have been given to the old single-race inquiry. Two broad categories of methods are identified: (1) *whole assignment* assigns individuals to a single-race category and (2) *fractional assignment* assigns individuals to multiple races on a fractional basis. We prefer whole assignment. Fractional assignment may be adequate for descriptive measures, but it would create complications for most analyses that cross census years. Moreover, regardless of analytical technique, whole assignment will be considerably simpler to use.

It is premature to articulate specific strategies for allocating Census 2000 race responses to create a backward-compatible race variable. Demographers have offered multiple suggestions for bridging race statistics in Census 2000 to earlier data (e.g., Del Pinal et al. 2001; Farley 2000; Goldstein and Morning 2000, 2002; Jones and Smith 2002; Lee 2001; Morning 2000, 2002; Office of Management and Budget 2001). As Lee (2001) notes, however, designing valid bridge variables will be impossible until we have individual-level data from Census 2000 to compare with other sources. We expect to adopt a probabilistic approach that assigns a primary race code to multiracial individuals depending on their individual characteristics and geographic location. Such an allocation approach will be based in part on survey data that ask both a multirace question and a primary-race question. Probably the National Health Interview Survey (NHIS) provides the best survey data for this purpose. Since 1976, the NHIS has allowed respondents to choose more than one racial category, and when respondents identified more than one racial group, they were asked the follow-up question: "Which of those groups would you say best describes your race?" In 2000, the response categories in the NHIS race questions were modified to conform to the five basic OMB categories: American Indian or Alaskan Native; Asian; black or African American; Native Hawaiian or Other Pacific Islander; and white.

Occupation, Industry, and Socioeconomic Indices

The coding of occupation and industry in Census 2000 is challenging. The Standard Occupational Classification system (SOC-1998) has undergone the most dramatic revision since 1940 (Bureau of Labor Statistics 1999). Unlike previous revisions, which largely added detail in some areas and removed it from others, the new system discarded all previous classifications and rearranged the entire structure with little concern for historical compatibility. To address disclosure concerns, the 5 percent census microdata sample will also combine many of the SOC-1998 categories into broader occupation groups. The new industrial classification system, the North American Industry Classification System (NAICS), was developed to allow full comparability of the United States, Canada, and Mexico. Like the new occupational classification, NAICS represents a major revision of previous standards. To accommodate the new occupation and industry classifications, we plan to develop new approaches.

IPUMS occupation coding strategy. Occupation is the most complex individual-level variable collected by the census. It is also among the most important census variables; excluding the basic demographic variables (age, sex, and race), occupation is the most frequently used variable in published IPUMS research. One of the key contributions of the IPUMS has been the reconciliation of occupation coding over the entire period from 1850 through 1990 (Sobek 1991, 1995, 1996, 2001; Sobek and Dillon 1995; Ronnander 1999). This reconciliation was based on the 1950 Census Bureau classification system because it posed the fewest technical difficulties and was familiar to social scientists.

For the period before 1940, occupations are coded directly into the 1950 system from the original open-ended response, so classification error is minimized. For the period since 1940, however, the IPUMS 1950 occupational classification is imperfect. From 1950 onward, changes in the Census Bureau occupational classification were mainly incremental, and most of the changes involved adding further detail. In each year, however, the Census Bureau moved specific occupational titles between categories. The Census Bureau has provided sufficient technical information to determine that many people were shifted among classifications between each census (U.S. Census Bureau 1968, 1972b, 1989). By working backward from the most recent census year, we identified which 1950 category would have contained a plurality of respondents for each category in each census year. The resulting standardized classification, though not flawless, has proven serviceable.

For Census 2000, our former strategy is no longer satisfactory. The SOC-1998 classification represents the most radical departure from previous practice since 1940. The imposition of 1950 occupation codes will therefore have to bridge 50 years and a previous substantial reclassification of occupations in 1980. If the Census Bureau provides an analysis of the statistical relationship between particular categories in the 1990 and 2000 classification systems, we will follow the aforementioned method to create 1950 occupation codes, which would be intended primarily as a blunt tool for long-term historical comparisons with the present. The resulting classification would be highly tenuous for analyses focusing on recent decades.[1]

Because of the basic incompatibility of Census 2000 with earlier census years, we plan a new approach to the harmonization of occupations. Instead of forcing all the classifications into the 1950 system, we will identify a lowest common denominator of occupations that can be consistently identified in all census years. To retain as many categories as possible, we will allow a small degree of misclassification and fully document where they occur. We expect that we will be able to identify between 50 and 100 occupation groups with reasonable consistency across the entire period from 1940 to 2000. We will then develop period-specific detail codes that will allow researchers to consistently identify larger numbers of occupations over briefer time spans. The original occupational classifications for all census years will also be available.

Industry coding. The industry codes are less problematic than the occupation codes. Although the new system, the 1997 NAICS, represents a major shift in industrial classification, our preliminary analysis suggests that the changes are more often related to the overall organization of the classification than to redefinition of particular categories. To maximize comparability with current data, we will create a simplified version of the NAICS that can be consistently identified in all census years. We will then add at least one additional digit to provide industrial detail available for only a subset of years.

Socioeconomic indicators. The IPUMS presently provides two socioeconomic indicators based on occupation. The first of these is the occupation score, which represents the median income for each occupational title in 1950. The second measure is the Duncan Socioeconomic Index (SEI), which is also pegged to the 1950 coding system (Duncan 1961). These measures are widely used, and we want to create similar variables that can be effectively applied to all census years. We will not confine ourselves to replicating the existing IPUMS measures (Sobek 1995); we expect to base the new SEIs on more recent measures of occupational incomes and prestige (e.g., Stevens and Cho 1985; Nakao and Treas 1992; Nam and Terrie 1988; Ganzeboom and Treiman 1996; Hauser and Warren 1997). We will develop these measures in consultation with experts in the field and with data users.

Geographic Harmonization of Public Microdata

To maximize the comparability of geographic variables, we plan to distinguish all areas that can be consistently identified across microdata samples from 1940 to 2000. This improvement in harmonization of census geography is feasible because of the National Historical Geographic Information System (NHGIS), a new NSF social science infrastructure initiative (Fitch and Ruggles, pp. 41–51 in Part One of this issue). The NHGIS is reconciling historical census cartography and creating historical electronic boundary files for census tracts, counties, and other geographic units.

Public use microdata area (PUMA) boundaries are identified in the Census Bureau electronic boundary files (TIGER) for 1990 and 2000. We will build boundary files for the geographic units identified in the microdata of earlier census years by aggregating the tract and county boundaries developed by NHGIS. The resulting database will enable IPUMS users to map public census microdata. Just as important, it will allow us to distinguish places that can be consistently identified across census years.

Inconsistencies in the coding of geographic areas have long posed frustrating compatibility issues for public microdata users. For the period from 1950 to 2000, the Census Bureau altered the definitions of substate geographic indicators in every census year; in 1980 and 1990, different geographic coding systems were used for the 5 percent and 1 percent microdata files. Although the IPUMS already contains some variables that attempt to harmonize census geography, by taking advantage of the NHGIS boundary files we can achieve significantly greater harmonization.

We will build compatible geographic variables on a state-by-state basis, proceeding backward from Census 2000. We will begin by overlaying the PUMA boundaries for 1990 and 2000 and identifying all areas that can be consistently delineated in both census years according to either the 1 percent or the 5 percent file boundaries. We will identify near matches—defined as geographic areas with less than 2.5 percent mismatch with respect to population coverage—by using aggregate-level census tract data. We will then examine nonmatching PUMAs manually to see if merging multiple areas will allow the identification of additional consistent areas. We will then repeat the process for each earlier census year. The areas will be numbered according to the standard IPUMS composite coding system so that users will be able to identify the geographic units common to any combination of census years. When feasible, we will construct compatible variables for the pre-1940 period, as well as the recent census years, so that users can make consistent long-run comparisons.

Restricted-Use Historical Census Files

The National Historical Census Files Project (NHCS) is a bold initiative of the Census Bureau to recover, preserve, document, harmonize, and disseminate all surviving machine-readable population census microdata from 1940 through 1990 (Gardner 2001). The NHCS project has designated the IPUMS coding system as a standard format for the data. These files include substantially greater geographic and subject area detail than is available in public use census microdata. The Census Bureau will make these data available to researchers only through Census Bureau Research Data Centers, which provide the necessary security for confidential data. These harmonized restricted files—in conjunction with an expanded and improved IPUMS—will allow social scientists to address simultaneously the broad sweep of time and the detail of spatial organization.

Table 2 summarizes the files contained in the restricted-use data series. These files begin in 1940 because older data are no longer subject to disclosure rules. The 1940 and 1950 restricted data files will be identical to the public use files, except that they will have detailed geographic identifiers: the public use files for those census years identify state economic areas (SEAs), whereas the restricted files will provide geographic identification down to the level of the enumeration district.

The large restricted-use census files begin in 1960. Unfortunately, the electronic complete-count short-form data for 1960 have been lost, but all long-form data survive.[2] There were two long-form questionnaires: one was distributed to 20 percent of households and the other to 5 percent. The two questionnaires were similar, differing only for a few housing items. For most purposes, the two samples can be combined to provide information on 25 percent of the population. This number of cases—some 44 million individuals—is sufficient to carry out analyses at the tract level.

For 1970 through 2000, the Census Bureau has preserved both short-form and long-form microdata. The 1970 census had two significantly different long-form question-

TABLE 2. Principal Data Files in the Restricted Microdata Series

			No. of records (in thousands)			IPUMS
Census year	Description	Sample density	Household	Person	IPUMS variables	file size (megabytes)
Restricted files						
1940	General sample	.01	391	1,351	179	584
1950	General sample	.01	461	1,922	175	798
1960	5% long form	.05	2,895	8,900	148	3,951
1960	20% long form	.20	11,580	35,600	148	15,805
1970	5% long form	.05	3,720	10,150	213	4,645
1970	15% long form	.15	11,160	30,450	213	13,935
1970	Short form	.80	59,520	162,400	32	11,148
1980	Long form	.19	17,898	43,073	283	20,425
1980	Short form	.81	76,302	183,627	32	9,846
1990	Long form	.16	17,696	40,000	259	19,328
1990	Short form	.84	92,904	210,000	30	11,754
2000	Long form	.16	18,112	45,024	259	21,645
2000	Short form	.84	95,088	236,376	28	12,626
2003	ACS	.03	3,429	8,526	259	4,005
2004	ACS	.03	3,462	8,610	259	4,044
2005	ACS	.03	3,498	8,697	259	4,085
Total			418,116	1,034,706		158,625

Note. ACS = American Community Survey.

naires, which went to 15 percent and 5 percent of the population, respectively. Since 1980, there has been just one long form, answered by approximately 19 percent of the population in 1980 and 16 percent of the population in 1990 and 2000. The absolute number of long-form respondents has remained nearly constant over the entire period from 1960 to 2000 because the rising population has compensated for the decline in sampling fraction over the past four decades. Under current plans, the public ACS microdata will be limited to a 1 percent annual sample, but the restricted ACS files will include the full 3 percent annual sample. In toto, the restricted data archive represents 16 times the number of records contained in the existing IPUMS for 1960 through 2000.

With the extraordinary demands of current data releases, the Census Bureau cannot devote substantial resources to the project. With the contribution of effort on the part of the University of Maryland Center on Population, Gender, and Social Inequality, the Census Bureau has sufficient resources to produce a clean and complete ASCII-format version of the data and to identify differences between the restricted data and the IPUMS coding system. This work involves four basic tasks: (1) transferring all raw files for both long-form and short-form data from a variety of obsolete media and formats to modern tape formats and converting them to column-delimited compressed ASCII files; (2) checking the converted files for completeness by comparing population totals for geographic areas with published totals; (3) verifying data conversion through new national-level tabulations of marginal frequencies for both original format and converted format data; and (4) comparing the converted restricted files with the public use versions of the data to identify incompatibilities with IPUMS coding designs.

The IPUMS Redesign project will build on this work and provide the tools that the Census Bureau needs to create, maintain, and disseminate an IPUMS-format version of the restricted data with harmonized geography, full documentation, and efficient data access tools. This work involves six additional tasks: (1) redesigning IPUMS variable coding to incorporate the full detail in the underlying data in all census years and to accommodate Census 2000 changes; (2) developing IPUMS-format data translation tables that describe all needed data transformations; (3) producing data conversion software to reformat and recode the data and to create standardized constructed variables; (4) designing and implementing a set of harmonized geographic variables that will allow consistent analysis of places across time; (5) expanding IPUMS documentation to cover differences between the public files and the restricted files; and (6) modifying the IPUMS data access system for use by the Census Bureau.

Public Test Files for the Restricted Microdata

The close similarity of the public data and the Census Bureau restricted files will allow cost-effective research in the Research Data Centers. There are presently eight centers located in Washington, Boston, Pittsburgh, Durham, California (Berkeley and Los Angles), Chicago, and Ann Arbor. Researchers wishing to use the centers must apply to the Census Bureau, and proposals must not only describe precisely the specific analyses to be undertaken but also explain how the proposed research contributes to the bureau's mission.

The centers are expensive; basic seat charges run $3,125 monthly, and many researchers also have to pay for travel, housing, and ancillary charges. Thus, major projects involving substantial data manipulation—such as merging confidential and public data sources—require considerable investment by funding agencies (Ruggles 2000). Even less-sophisticated analyses can be expensive because users cannot properly debug programs before entering the center.

We plan to make it possible for researchers to develop and test their analyses and data transformations using freely available public data designed to simulate the restricted files. By demonstrating the feasibility of their proposed methods, researchers will speed up the approval process for their applications to use the centers. More important, because fully tested software programs (or statistical software command files) are likely to work on the first attempt, less time will be needed within the centers. By reducing costs and freeing space within the centers, the availability of test data will make these important resources accessible to a broader range of researchers.

Preparing suitable test data will require some alteration of the public microdata. Researchers mainly need access to the restricted files because they include more information than the public files. The most important of these variables are the low-level geographic identifiers. The restricted files also contain more detail for other variables. For example, the restricted files do not top-code variables such as income or age, and they provide more information about race and institution type.

We will therefore develop public test files that simulate the variables and categories in the restricted data. We will begin with the public versions of the microdata and then add simulated low-level geographic variables derived from the geographic distribution within microdata areas of race, age, sex, and household type. On the basis of their characteristics, households will be probabilistically assigned to census tracts. This simulated tract data will then be used to construct all the higher-level restricted-file geographic variables, such as county, city, and minor civil division.

We will modify the nongeographic variables by using similar methods. On the basis of a theoretical income distribution, we will randomly assign high incomes to top-coded individuals, and we will assign detailed institution types randomly on the basis of the marginal distributions. Analyses based on the imputed variables will, of course. yield incorrect results. However, because all variables and values will be present and their distributions plausible, the resulting data set will be suitable for test purposes.

We will distribute the test data set from a Web site designed specifically for the restricted data files. The data access system on this site will mimic the look and feel of the Census Bureau internal site, so users will know what to expect. We will also provide information about the Research Data Centers, advice on preparing research proposals, and samples of successful applications.

Harmonized Geography for Restricted Census Files

The restricted historical census files for 1940 and 1950 identify state, county, metropolitan area, minor civil division, and enumeration district. From 1960 onward, they also distinguish wards, census tracts, and census places, and from 1990 onward they include census blocks and block groups. In all cases, these geographic units are defined according to contemporary standards, so the boundaries vary from one census to the next. Using the NHGIS boundary files, we will design geographic coding systems that identify the same geographic areas in all census years by means of a standardized numbering system. These variables will include tracts and tract groups, counties, cities, and PUMAs defined according to Census 2000 boundaries.

The procedure for developing the new geographic codes is straightforward. Suppose, for example, we want to impose Census 2000 PUMA boundaries on all earlier census years. Using the NHGIS database, we can automatically generate a list of census tracts in each prior census year coded according to contemporary standards that fall entirely within each Census 2000 PUMA. A small percentage of tracts, particularly in earlier census years, will cross PUMA boundaries. We will handle these situations on a case-by-case basis. When most of a tract falls within one PUMA, we will allocate the entire tract to that PUMA if it would yield less than a 2.5 percent error in the PUMA population. If the error is greater, we will subdivide the tract into block groups or enumeration districts and allocate part of the tract to each PUMA. The result will be an equivalency file that can be run against the restricted files to add a consistent 2000 PUMA identifier.[3]

We will define the types of geographic areas in collaboration with Census Bureau staff, users, and experts in the field. We anticipate including at least one low-level identifier at the approximate level of census tracts. In addition, we will construct standardized boundaries for various administrative and statistical units, such as counties, municipalities, and metropolitan areas. We will also investigate the potential for developing consistently defined labor-market areas.

Potential for New Public Files

The NHCS project is extraordinarily valuable. The large size and fine geographic detail of the restricted files will allow a wide range of new multilevel analyses and studies of small population subgroups. The studies also open the door to new public use microdata samples of 1960 and 1970 and improved aggregate data files for 1960 through 1980.

Currently, the 1960 census is the weak link in the IPUMS because it has no geographic codes below the state level and only a single 1 percent sample is available. Although they represent 6 percent of the population, the 1970 samples also use geographic identifiers that are not comparable with later census years. We will use the NHGIS cartographic database to design new geographic systems for 1960 and 1970 that are compatible with current Census Bureau disclosure standards. These systems will make it easy for the bureau to produce a new 5 percent sample for 1960 and new versions of existing 1960 and 1970 samples with enhanced census geography.

We envision 1960 as a bridge year. The existing 1960 1 percent sample identifies no places below the state level. We plan to match the records in the existing 1960 sample to the restricted data file so that we can add a variable identifying SEAs, the smallest geographic unit available in the 1940 and 1950 samples. SEAs are also identified for the censuses of 1850 through 1930, so the new variable will provide a fully consistent substate geographic indicator spanning 110 years. In addition, we will create geographic identifiers for a new 5 percent file of 1960 that maximize comparability with the census county group and PUMA classifications for 1980, 1990, and 2000.

For 1970, we will direct our efforts toward creating improved geographic identifiers for the existing data sets.[4] The 1970 samples are currently divided into three file types with differing geographic coding systems: state files, county group files, and neighborhood files. At present, the state files have no geographic information below the state level, so we will create new geographic identifiers that maximize compatibility with recent census years, just as we propose to do for the new 1960 5 percent sample. The county groups—which currently identify 402 areas that had between 250,000 and 500,000 population in 1970—are less flexible. Nevertheless, they also have potential for improvement because these existing county groups identify considerably less geographic detail than is allowed under current disclosure rules. We will subdivide the existing county groups to design a new system that identifies approximately 1,400 areas.

In addition to new public microdata products, the restricted census files could serve as the source for creating improved public tabular data describing small areas for the period 1960 through 1980 that meet current standards with respect to subject coverage, detail, geographic coding, and confidentiality. New summary files would greatly improve comparability of aggregate census data over time. For example, the 1960 aggregate census files provide only two race categories, white and nonwhite, even though the restricted microdata files from 1960 include sufficient detail to replicate the principal race categories tabulated in recent censuses before 2000.

At this writing, we have no guarantee that the Census Bureau will release new historical public use microdata or summary files; that decision rests with the review board. Our work, however, will make it simple and cost-effective for the bureau to produce new historical files that meet current standards of disclosure limitation and documentation. Considering the strong interest of the research community, we are confident that the review board will consider releasing the files to the public.

Metadata and Software

To accommodate the dramatic technological changes of the past decade and ensure the long-run sustainability of the IPUMS, we need to bring the database infrastructure up to date. We plan four major initiatives: (1) revising the documentation to conform to new eXtensible Markup Language (XML) metadata standards; (2) implementing a new system for version control; (3) rewriting all software used to create and disseminate the IPUMS; and (4) mirroring the entire IPUMS system at the Inter-university Consortium for Political and Social Research (ICPSR).

Machine-Understandable Metadata

We have retired the printed version of the IPUMS documentation, and it is now stored in approximately 2,800 Web pages. Most of these are static pages, but an increasing number are dynamic pages constructed automatically when users request them. This arrangement has many advantages, but it also creates three problems: (1) long-term preservation is a concern because the documentation is system specific and hardware dependent; (2) the continuous process of editing and correcting individual Web pages creates serious issues of documentation version control; and (3) the system is difficult to maintain.

To address these problems we will convert the IPUMS documentation into machine-understandable metadata and will adopt the Data Documentation Initiative (DDI) metadata standard. As described elsewhere (see the article by Block and Thomas in Part Two of this issue), the DDI is a nonproprietary, hardware-independent, neutral documentation standard. As a product of the world's leading data archives, the DDI's primary goal was to establish an archival standard for documentation to reduce the costs of long-term preservation and access. Thus, the system addresses our concerns about documentation sustainability. Most important, perhaps, the DDI will reduce the costs of system maintenance and decrease the potential for documentation errors. In a DDI codebook, each item is tagged with information about its meaning; thus, the codebook has a machine-understandable structure that allows for automated processing by data access software.

We intend to modify the IPUMS data and documentation access system so that it is driven by DDI-compliant metadata. Once the new system is in place, it will be possible to modify a variable by changing its specifications in a single location. The software will then propagate that change throughout the system. This approach will increase the flexibility of the IPUMS and greatly simplify the incorporation of new data files into the system.

Version-Control Protocols

The ability to replicate existing studies is essential to the scientific enterprise; it provides our fundamental means of understanding, evaluating, and building upon past research. The IPUMS needs a rigorous system for version control to ensure that scholars can replicate past IPUMS-based results.

Since the first general release of the IPUMS in November 1995, we have made several hundred improvements to the data and thousands of updates of the documentation. A shortcoming of the project has been our failure to impose thorough version control. In general, we have corrected errors whenever we have uncovered them and added improvements whenever time and funding have permitted. Since 1998, we have made an effort to note major changes on the revisions page of the IPUMS Web site (http://www.ipums.org/usa/revisions.html), but the effort has not been comprehensive. Users have no means of determining which version of the data they are using; furthermore, when we change the data, users have no means of restoring the version they used in a previous analysis.

Version control is therefore a crucial issue. With the expansion and extensive revision of the data series, it is now more critical than ever to develop a comprehensive solution. With the dramatic decline in data storage costs, retaining old versions of the IPUMS data is now feasible. We will design and implement a system that will clearly identify, within both data and documentation, the precise version of the IPUMS that was used to create an extract. In addition, the system will allow anyone to replicate any past data extract from an earlier IPUMS version and to identify the specific changes that occurred between any two versions.

We will add version control to the component of the IPUMS data extract system that currently allows users to replicate or modify past extracts with the most current version of the data. When users create an extract using the current system, they receive a customized short codebook for the data file. In the future, that codebook will contain version numbers for the data and the documentation. It will also include a recommended citation incorporating a unique number for a particular extract. We will add a feature to the extract system that will allow anyone to specify an extract number and obtain a replica of that extract based on the same data version. Thus, if scholars identify an extract number in their publications, readers of their work will be able to create and download an exact copy of the data used for the research.

Software

IPUMS Redesign will require significant software development. The existing IPUMS software falls into two main categories: (1) the data conversion programs recode and reformat the data; check for internal consistency; detect and diagnose design flaws; and create new constructed variables on poverty, socioeconomic status, geography, family interrelationships, and other subject areas; and (2) the data access programs control the Web-based interface between data and documentation, maintain user accounts, and create customized subsets of IPUMS data.

Data conversion software. The existing IPUMS data conversion programs are aging and inadequate. They consist of approximately 18,000 lines of FORTRAN code written mainly between 1991 and 1993 and continuously modified since. The software was designed to handle only the nine census years that existed when it was written. It was not conceived with a flexible structure, so the addition of new census years has required a series of makeshift extensions. Moreover, the programs are poorly documented, awkwardly structured, and difficult to maintain.

To accommodate Census 2000, the ACS, and the restricted Census Bureau files, we must extend the data conversion software to handle these new data files. To convert the restricted files in a secure environment, the software must be installed at the Census Bureau; therefore, the programs must be transparent enough to be understood and modified by bureau personnel.

It is not cost-effective to modify the existing programs to meet these goals; the structural problems are great enough that it is cheaper to build a new house than to renovate the old one. We are developing a flexible and expandable system under open-source standards. In addition to serving the immediate needs of the Census Bureau, understandable and flexible open-source software will serve other important goals.

The software used to create the data is the ultimate documentation. Although we have invested considerable effort to describe our procedures clearly, the English language lacks the necessary precision to convey data transformation procedures unambiguously. Accordingly, the ICPSR has agreed to archive the IPUMS software and data translation tables as well as the data and documentation, so future researchers can reconstruct exactly what we did.

Data access software. The IPUMS programs for Web-based access to data and documentation are of much more recent vintage than the data conversion software. During the past two years, the professional programming staff has redesigned and thoroughly documented all data access systems. Nevertheless, we plan further enhancements of the data access system that take advantage of metadata development and other new Minnesota Population Center (MPC) projects.

The adoption of DDI-encoded metadata has the potential to greatly simplify the maintenance of documentation, but only if the documentation browser and data extraction system are driven by the DDI codebook. Most IPUMS documentation is currently stored in static HTML pages; the rest of the software is driven by a series of tables describing the content of the data. XML is well suited to Web-based applications, but our use of DDI-compliant metadata nevertheless represents a fundamental shift in strategy and will require careful redesign of data and documentation access software. We will also be adding several new features to the data extraction system, such as the version control protocols.

We also plan new software that will allow IPUMS users to capitalize on the NHGIS. In addition to improving geographic compatibility across census years, the NHGIS will open new opportunities for multilevel analysis of census microdata. We plan to develop an interface between the IPUMS and the NHGIS that will allow researchers to attach aggregate statistics to individual-level records as part of an IPUMS data extract. These statistics will describe PUMAs, county groups, SEAs, cities, counties, metropolitan areas, states, or other geographic units identified in a microdata sample. By automating the tedious process of merging microdata and aggregate data, we hope to enrich both sources and stimulate a broad range of new research initiatives.

Preservation and Sustainability

We have already described two key innovations that will promote long-run preservation and sustainability of IPUMS data. First, DDI-encoded metadata will ensure that the IPUMS remains usable even if the technological environment shifts dramatically. Second, redesigned open-source software will allow the Census Bureau to maintain and expand the restricted files indefinitely.

We are also concerned about long-term maintenance of the public use IPUMS database. The MPC is not a permanent data archive; future funding is uncertain, and we must prepare for the possibility that it may disappear. Accordingly, we are collaborating with the ICPSR to develop and implement a permanent plan for preservation of IPUMS data and software. Archiving the IPUMS is especially complicated for the following reasons: the scale and complexity of the database, the need to install, preserve, and maintain software as well as data, and the desirability of archiving multiple versions of the data. Our collaboration with the ICPSR has two principal goals. First, as soon as possible we will establish the ICPSR as the permanent repository for the principal IPUMS data versions from 1995 onward. In effect, this means that there will be a backup copy of the data in the form of hierarchical ASCII files with DDI-compliant documentation and SAS, SPSS, and Stata data definition statements. Second, we will establish the ICPSR as a mirror for the IPUMS Web site. This collaboration will provide both immediate security and assurance of long-term support for the data access system.

Discussion

Redesign of the IPUMS is needed for several reasons. The existing IPUMS design was based on Census Bureau standards for the 1980 and 1990 microdata, and it is not optimal for Census 2000 or the ACS. Moreover, the technological environment has shifted significantly since the IPUMS was created, and we need to make modifications in both variable design and software to capitalize on these changes. In addition, some improvements to the IPUMS—such as the geographic coding—should have been incorporated from the outset, but they have been delayed owing to limited resources.

The incorporation of Census 2000 and the ACS into the IPUMS will link the historical data series to the present, allowing researchers to analyze current policy issues in historical perspective. Merging recent and historical data creates a powerful research tool, one that is not just for historians. Models and descriptions of the past underlie both theories of past social change and projections into the future. The series of IPUMS samples provides a unique laboratory for the study of economic and demographic processes. This kind of empirical foundation is essential for developing and testing social and economic models. On the basis of the strong interest in Census 2000 and the ACS that our users have already expressed, we expect the addition of new data to generate a substantial surge in new IPUMS-based research.

Just as exciting is the prospect of access to the internal restricted microdata files for the period 1960 to 2000. These data represent a new class of source material for social scientists. We anticipate that the availability of consistent microdata for the entire population over a broad time span will have a profound effect on the practice of social science research, comparable in its impact to the first release of census microdata in 1964. Among key substantive areas are residential segregation, the decline and renaissance of central cities, immigrant and ethnic settlement patterns, suburbanization and urban sprawl, rural depopulation and agricultural consolidation, the identification of concentrated poverty, transportation, and public health and epidemiological studies.

The potential for spatial population analysis is particularly promising. In virtually every subject area—including education, poverty, and basic demography—spatial analysis has been handicapped by the limitations of crude aggregate-level statistics available for small areas. Access to the individual-level data is important because it will not only allow researchers to control for individual and household characteristics but will also allow researchers to tabulate customized aggregate-level measures tailored to particular research questions. This access will open new opportunities for multilevel analysis that simultaneously considers the characteristics of individuals, households, census tracts, labor markets, metropolitan areas, states, and regions. Most important, because the data incorporate multiple census years, they will allow social scientists for the first time to address simultaneously the broad sweep of time and the detail of spatial organization.

NOTES

[1] Matters are further complicated because in the 5 percent public microdata sample of Census 2000, a simplified version of SOC-1998 will be used to eliminate any occupational categories with fewer than ten thousand persons in the general population. As many as 100 of the 600 titles in the system may be eliminated; we will not know the final outcome until the Census Bureau has analyzed the data and determined a strategy for collapsing categories.

[2] The 1960 short forms do survive on microfilm, and they were designed for optical scanning by means of the Film Optical Sensing Device for Input to Computers (FOSDIC). Some discussion has taken place within the Census Bureau about recovering the 1960 short-form data by rescanning the film, an expensive project that is unlikely to be undertaken inside the time frame of this project. We will nevertheless design the system to anticipate the future addition of 1960 short-form data.

[3] For 1940 and 1950, there is an additional step. The restricted files do not contain tract identifiers, and the NHGIS database does not include enumeration district information. Therefore, to harmonize geography for these census years, we would have to develop equivalency files between enumeration districts and tracts. There were some 175,000 districts in those two census years, so it is not a trivial task. We are now investigating the feasibility of creating such files.

[4] The Census Bureau cannot release a new sample of the 1970 public use files because the Census Bureau's Disclosure Review Board has adopted a policy that information on no more than 6 percent of the population will be publicly released in the form of microdata, and 6 percent of the 1970 census is already in the public domain.

REFERENCES

Bureau of Labor Statistics. 1999. Revising the standard occupational classification system. Report 929. Washington, D.C.: GPO.

Del Pinal, J. H., et al. 2001. Reporting of two or more races in the 1999 American community survey. Working Paper No. 329, May. Annandale-on-Hudson, N.Y.: Jerome Levy Economics Institute, Bard College.

Duncan, O. D. 1961. A socioeconomic index for all occupations. In *Occupations and social status*, edited by A. Reiss. New York: Russell Sage.

Farley, R. 2000. Racial identities in the Census of 2000. What did respondents do when they had the opportunity to identify with multiple races? Paper presented at the Jerome Levy Economics Institute, Bard College, Conference on Multiraciality: How Will the New Census Data Be Used? 22–23 September.

Ganzeboom, H., and D. Treiman. 1996. Internationally comparable measures of occupational status for the 1988 international standard classification of occupations. *Social Science Research* 25: 201–39.

Gardner, T. 2001. The National Historical Census Files Project. Paper presented at the biennial Conference of Official Representatives of the Inter-university Consortium for Political and Social Research, Ann Arbor, October.

Goldstein, J., and A. Morning. 2000. The multiple-race population of the United States: Issues and estimates. *Proceedings of the National Academy of Sciences* 97: 6230–35.

———. 2002. Back in the box: Allocating multiple-race responses back to single races. In *The new race question: How the census counts multiracial individuals*, edited by J. Perlmann and M. Waters, 119–36. New York: Russell Sage Foundation.

Hauser, R. M., and J. R. Warren. 1997. Socioeconomic indexes for occupations: A review, update, and critique. *Sociological Methodology* 27: 177–298.

Jones, N. A., and A. S. Smith. 2002. Who is multiracial? Exploring the complexities and challenges associated with identifying the multiracial population in Census 2000. Paper presented at the annual meeting of the Population Association of America, Atlanta, May.

Lee, S. M. 2001. Using the new racial categories in the 2000 census. Kids Count/Population Reference Bureau Report. Washington, D.C.: Annie E. Casey Foundation and Population Reference Bureau.

Morning, A. 2000. Who is multiracial? Definitions and decisions. *Sociological Imagination* 37: 209–29.

———. 2002. New faces, old faces: Multiracial enumeration in the United States, past and present. In *New faces in a changing America: Multiracial identity in the 21st century,* edited by H. DeBose and L. Winters, 41–67. Thousand Oaks, Calif.: Sage.

Nakao, K., and J. Treas. 1992. The 1989 socioeconomic index of occupations: Construction from the 1989 occupational prestige scores. GSS Methodological Report No. 74. Chicago: National Opinion Research Center.

Nam, C., and E. W. Terrie. 1988. 1980-based Nam-Powers occupational status scores. Working Paper Series 88–48. Tallahassee: Center for the Study of Population, Florida State University.

Office of Management and Budget. 2001. Provisional guidance on the implementation of the 1997 standards for federal data on race and ethnicity. Washington, D.C.: OMB Federal Register, 16 January.

Ronnander, C. 1999. The classification of work: Applying 1950 census occupation and industry codes to 1920 responses. *Historical Methods* 32: 151–55.

Ruggles, S. 2000. A data user's perspective on confidentiality. *Of Significance . . . a Topical Journal of the Association of Public Data Users* 2: 1–5.

Sobek, M. 1991. Class analysis and the U.S. census public use samples. *Historical Methods* 24: 171–81.

———. 1995. The comparability of occupations and the generation of income scores. *Historical Methods* 28: 47–51.

———. 1996. Work, status, and income: Men in the American occupational structure since the late nineteenth century. *Social Science History* 20: 169–207.

———. 2001. New statistics on the U.S. labor force, 1850–1990. *Historical Methods* 34: 71–87.

Sobek, M., and L. Dillon. 1995. Occupational coding. *Historical Methods* 28: 70–73.

Stevens, G., and J. H. Cho. 1985. Socioeconomic indexes and the new 1980 census occupational classification scheme. *Social Science Research* 14: 142–68.

U.S. Census Bureau. 1964. *Census of population and housing, 1960 public use sample: One-in-one thousand sample.* Washington, D.C.: GPO.

———. 1968. Changes between the 1950 and 1960 occupation and industry classifications, by J. A. Priebe. Technical Paper 18. Washington, D.C.: GPO.

———. 1972. 1970 occupation and industry classifications in terms of their 1960 occupation and industry elements, by J. A. Priebe, J. Heinkel, and S. Greene. Technical Paper 26. Washington, D.C.: GPO.

———. 1989. The relationship between the 1970 and 1980 industry and occupation classification systems, by P. Vines and J. A. Priebe. Technical Paper 59. Washington, D.C.: GPO.