Disseminating Historical Census Data on the World Wide Web

by Steven Ruggles, Matthew Sobek, and Todd Gardner[1] *,University of Minnesota, Department of History*

Introduction

This paper describes our project to electronically disseminate the Integrated Public Use Microdata Series (IPUMS). The IPUMS—the world's largest publicly available demographic database—is a coherent series of individual-level census data drawn from eleven census years between 1850 and 1990. Prior to the IPUMS, the U.S. census samples were a haphazard assortment of files created by different researchers at different times, each with its own unique record layout and coding schemes. By putting all the census samples in a compatible format, the IPUMS greatly simplifies the use of multiple census years. The database constitutes our most powerful resource for the study of long-term American social and economic change.

The development of the Internet and World Wide Web (WWW) promises to transform fundamentally the nature of electronic data dissemination. At the same time, the proliferation of fast personal computers and UNIX workstations is already revolutionizing data analysis. Our project capitalizes on both of these developments by making the largest and most powerful population database readily available for analysis on desktop machines.

The scholarly community has already shown great interest in the IPUMS. Nevertheless, the sheer size of the database poses problems for researchers. The Internet provides the most practical means of dissemination, but current methods of data distribution on the Internet are limited. Although enormous computational resources are becoming available to researchers, access to large-scale microdata is still cumbersome and expensive. The logistical difficulties are compounded by the spotty to nonexistent support for any but the most recent census samples.

We are currently in the process of implementing a project that addresses these distribution and usability issues. It is composed of three complementary elements, the second of which we have not yet undertaken:

1. Development of a data extraction system for use on the WWW. The preliminary version, described in detail below, is already in place. Users can fashion subsamples containing only those years, subpopulations, and variables that suit their research interests and computing power. In the future, we envision a Java-based extract system that is truly interactive, incorporating a variety of advanced features enabling researchers to customize their data extracts. They will also be able to construct new variables that capitalize on the hierarchical structure of the database.

2. Conversion of the IPUMS documentation into hypertext format to facilitate navigation. Users will be able to jump instantly to relevant sections of the documentation with the click of a mouse without negotiating 3000 pages of text. The documentation will be integrated with the extract system so that researchers can make informed choices in designing their subsamples. By using Adobe Acrobat format, the documentation will be downloadable onto virtually any computer platform while retaining its hypertext functions.

3. Ongoing support for users. For the first time, all of the existing census samples will be supported, in their integrated format.

The IPUMS has the capacity to become a cornerstone in the infrastructure for social science research. Our project is intended to remove the remaining obstacles preventing a wide range of researchers from taking advantage of the unique resource that the census samples represent. We hope our project can serve as a model for the distribution of microdata more generally.

Context

Since 1978, the federal government has invested approximately $19 million (in 1995 dollars) to create historical Public Use Microdata Samples (PUMS) of the decennial censuses for the period 1850 to 1950 (Graham 1980; Strong 1989; Ruggles and Menard 1994; Ruggles et al 1995; U.S. Bureau of the Census 1984a, 1984b). Beginning with the 1960 census, the Census Bureau has produced PUMS as a byproduct of each decennial enumeration (U.S. Bureau of the Census 1972, 1973, 1982, 1992). We now have a series of microdata for eleven census years (1850, 1880, 1900, 1910, 1920, 1940, 1950, 1960, 1970, 1980 and 1990), and a proposal to create two more samples (for 1860 and 1870) was recently funded by NICHD.

Taken together, these data files comprise our most powerful resource for the study of historical social and economic change. The range of potential topics that can be addressed with the national census files includes household composition, fertility, life-course transitions, ethnicity, immigration, internal migration, female labor force participation, the household economy, industrial and occupational structure, urbanization, nuptiality, and education. High-precision sample designs allow national and regional estimates that are virtually as reliable as published census data, but which have far greater subject area coverage. As microdata, rather than aggregate summary data, the samples provide information about individual persons and households transcribed directly from the original census manuscripts. The microdata contain far richer information than was ever published in the census volumes. They enable researchers to make tabulations tailored to their specific research questions and to overcome incompatibilities in the published census tabulations. In addition, they have allowed researchers to move beyond simple tabular analysis and apply increasingly sophisticated multivariate techniques. Although most of these census files have only been available for a few years, they have already led to an outpouring of new research (e.g., Jacobs 1989; Hirschman and Kraly 1990; Mare 1991; Jenson 1991; Kalmijn 1994; Ruggles 1994a, 1994b; Watkins 1994; Gjerde and McCants 1995).

The national census files have three key strengths: complete geographic coverage, large sample populations, and broad chronological scope. Complete geographic coverage is important not only because it allows scholars to generalize at the national level; national samples can also provide context for local studies. Moreover, by linking the census microdata to aggregate sources describing local characteristics, the PUMS allow multi-level analyses of the effects of local conditions on individual and family behavior.

The second strength of the national public use census files is their large size. The number of cases available for each census year ranges from the hundreds of thousands to the tens of millions. This allows the study of small and geographically dispersed population subgroups. For example, researchers at the University of Minnesota using the historical public use samples have examined topics such as the professionalization of nursing, American Indian fertility patterns, the living arrangements of elderly urban blacks, the demography of the prison population, the gender composition of clerical workers, and the living arrangements of parentless children. These research topics could not be pursued using a general social survey of the scale ordinarily undertaken by academic social scientists. Indeed, even the largest social survey carried out by the government—the Current Population Survey—is far too small for the detailed analysis of topics such as American Indian fertility or the professionalization of nursing (Olson 1991; Shoemaker 1991). The public use samples are the only general source of microdata available for any period with sufficient cases to study such small population subgroups.

The third, and most important, strength of the historical public use census files is their potential for the study of social and economic change over long periods of time. There is no other consistent source of quantitative information about the American population spanning more than a few decades. Despite frequent changes in subject content and modifications of enumeration procedures, the core of the census has remained remarkably stable over the past century and a half (Magnuson 1995). Since 1850, every census consists of a listing of individuals within households in a prescribed sequence and provides data on basic demographic characteristics such as age, sex, race, and birthplace. Table 1 indicates the broad range of subject areas covered in each census year.

Although the PUMS files are the most widely used data in American social science, few researchers have exploited the great potential of the national census files for the study of change over time. Instead, most investigators use single samples as isolated cross-sections (e.g., Haines 1989; Johnson and Lean 1985; Sanderson 1987; Sandefur and Sakamoto 1988; Sorenson 1989; Gordon and McLanahan 1991; Morgan et al 1993; Farley and Frey 1994; Krivo 1995; Sassler 1995). This is mainly because of incompatibilities among the original samples. The PUMS were created at different times by different investigators and, as a result, they have incompatible documentation and a wide variety of record layouts and coding schemes. Several previous census microdata projects—the samples of 1900, 1910, 1940 and 1950—ran out of money before they were finished. In most cases, the basic data were fine, but compromises were made in dissemination, user support, and documentation. Indeed, in the cases of the 1940 and 1950 public use microdata samples, critical sections of documentation essential for the proper use of the samples were omitted.

To resolve the incompatibilities among census samples and limitations of their documentation, NSF funded a project entitled“Integrated Public Use Microdata Series.” The project incorporated all 23 existing national census samples into a single coherent database with an integrated set of documentation (Ruggles and Sobek 1995). The constituent samples are described in Table 2. All census years receive the same record layout and coding, without any loss of information from the original PUMS. Missing data in the early census years is imputed for the more important demographic variables. The

Table 1. Availability of Select Subject Areas Across Census Years

	185	1880	1900	1910	1920	1940	1950	1960	197	1980	1990
Household Record											
State	X	X	X	X	X	X	X	X	X	X	X
County	X	X	X	X	X	.	.	.	.	.	.
County group/public use microdata area	.	.	.	.	.	.	.	.	X	X	X
State economic area	X	X	X	X	X	X	X	.	.	.	.
Metropolitan status	X	X	X	X	X	X	X	X	X	X	X
Metropolitan area	X	X	X	X	X	X	X	.	X	X	X
City	X	X	X	X	X	X	X	.	.	X	X
Size of place	X	X	X	X	X	X	X	.	.	X	X
Urban/rural status	X	X	X	X	X	.	.	X	X	X	X
Farm	X	X	X	X	X	X	X	X	X	X	X
Ownership of dwelling	.	.	X	X	X	X	.	X	X	X	X
Mortgage status	.	.	X	X	X	.	.	.	.	X	X
Value of house or property	.	.	.	.	.	X	.	X	X	X	X
Monthly rent	.	.	.	.	.	X	.	X	X	X	X
Total family income	.	.	.	.	.	.	X	X	X	X	X
Person Record											
Relationship to household head	X	X	X	X	X	X	X	X	X	X	X
Age	X	X	X	X	X	X	X	X	X	X	X
Sex	X	X	X	X	X	X	X	X	X	X	X
Race	X	X	X	X	X	X	X	X	X	X	X
Marital status	.	X	X	X	X	X	X	X	X	X	X
Age at first marriage	.	.	.	.	.	X	.	X	X	X	.
Duration of marriage	.	.	X	X	.	.	X	.	.	.	.
Times married	.	.	.	X	.	X	X	X	X	X	.
Children ever born	.	.	X	X	.	X	X	X	X	X	X
Birthplace	X	X	X	X	X	X	X	X	X	X	X
Parents' birthplaces	.	X	X	X	X	X	X	X	X	.	.
Ancestry	.	.	.	.	.	.	.	.	.	X	X
Years in the United States	.	.	X	X	X	.	.	.	X	X	X
Mother tongue	.	.	.	X	X	X	.	X	X	.	.
Language spoken	.	.	.	X	.	.	.	.	.	X	X
School attendance	X	X	X	X	X	X	X	X	X	X	X
Educational attainment	.	.	.	.	.	X	X	X	X	X	X
Literacy	X	X	X	X	X	.	.	.	.	.	.
Employment status	.	.	.	X	.	X	X	X	X	X	X
Occupation	X	X	X	X	X	X	X	X	X	X	X
Industry	.	.	.	X	X	X	X	X	X	X	X
Class of worker	.	.	.	X	X	X	X	X	X	X	X
Weeks worked last year	.	.	.	.	.	X	X	X	X	X	X
Weeks unemployed	.	X	X	X	.	X	X	.	.	.	.
Total personal income	.	.	.	.	.	.	X	X	X	X	X
Wage and salary income	.	.	.	.	.	X	X	X	X	X	X
Migration status	.	.	.	.	.	X	X	X	X	X	X
Veteran status	.	.	.	X	.	X	X	X	X	X	X
Name	X	X	.	.	X	.	.	.	.	.	.

database also includes a series of fully compatible constructed variables. For example, the IPUMS provides “pointer” variables identifying the position within the household of every individual’s own mother, father, and spouse. These and other variables, constructed identically for all years, provide the building blocks for researchers to design their own variables. Documentation includes comparability discussions for every variable as well as separate essays on the more complicated aspects of the data series.

We began distributing a preliminary version of the IPUMS data in September 1993 through an anonymous FTP site, and in April 1995 added a World Wide Web site. User response to our early data products has been overwhelming. When we set up our site in 1993 to distribute beta-test copies of a preliminary version of the IPUMS, our goal was simply to obtain feedback from experienced researchers on the design of the database. We expected to distribute only a few copies of the data, but news of its availability on the Internet spread quickly through the research community. Figure 1 gives the volume of IPUMS data downloaded by outside researchers since the preliminary release in late 1993.

Interest in the IPUMS database among social scientists clearly is high. In March 1995, the University of California at Riverside sponsored an All-Campus University of California Economic History Conference that drew fifty current and prospective users of the IPUMS to discuss methodological issues and present early research results from the database. The population centers at the Universities of Michigan, Wisconsin, Texas, Minnesota, SUNY-Buffalo, and Chicago have invited us to describe the data and its applications to interested researchers and data archive staff. In addition we have carried out extensive correspondence with early users of the IPUMS data. We designed the project in response to input from all these sources.

Despite the intense interest in the IPUMS, many researchers still have problems managing such large files. The number of scholars who have made effective use of the data so far represent only a small minority of those potentially interested. Although a large number of researchers have accessed the data, after downloading a file or two many have realized they lacked the resources to manipulate and analyze the data. The IPUMS contains approximately 69 million records, spans 140 years, and incorporates 524 separate variables. The sheer size of the IPUMS database (25 gigabytes) presents new challenges for data distribution (see the last column of Table 2). The Internet provides the most practical means of dissemination, but current methods of Internet data distribution are limited. Storage and distribution issues are the most consistent complaints conveyed by researchers using the PUMS and IPUMS. Most users need to decompress the data in order to use it, but computers with sufficient storage to decompress even the 1-in-100 samples are rare. The development of distributed computing environments has placed enormous computational resources in the hands of researchers, but access to large-scale microdata is still cumbersome and expensive. Even the documentation for the IPUMS database suffers from size problems: at 3000 pages, it is almost as unwieldy as the data itself.

To the extent that poor documentation, inadequate dissemination, and limited user support have curtailed use of the microdata samples, the resources invested in the PUMS samples have been wasted. The IPUMS project has the potential to correct these deficiencies, but only if access and use can be simplified. The significance of our project is not limited to the support and dissemination of the IPUMS, however. We hope to provide a model for the distribution of microdata generally. Accordingly, we will make our extract program and hypertext interface available to all interested researchers.

Extract System Design

To address the aforementioned distribution and usability issues, we have developed an extract system for use on the WWW that allows researchers to select only those subpopulations and variables needed for a particular analysis. In the future, we will integrate the documentation with the extract system and present it in hypertext format to facilitate navigation. In addition, we will offer user support of all the existing census samples in their IPUMS format. Each of these three aspects of the project is described in detail below.

1. Interactive extract system

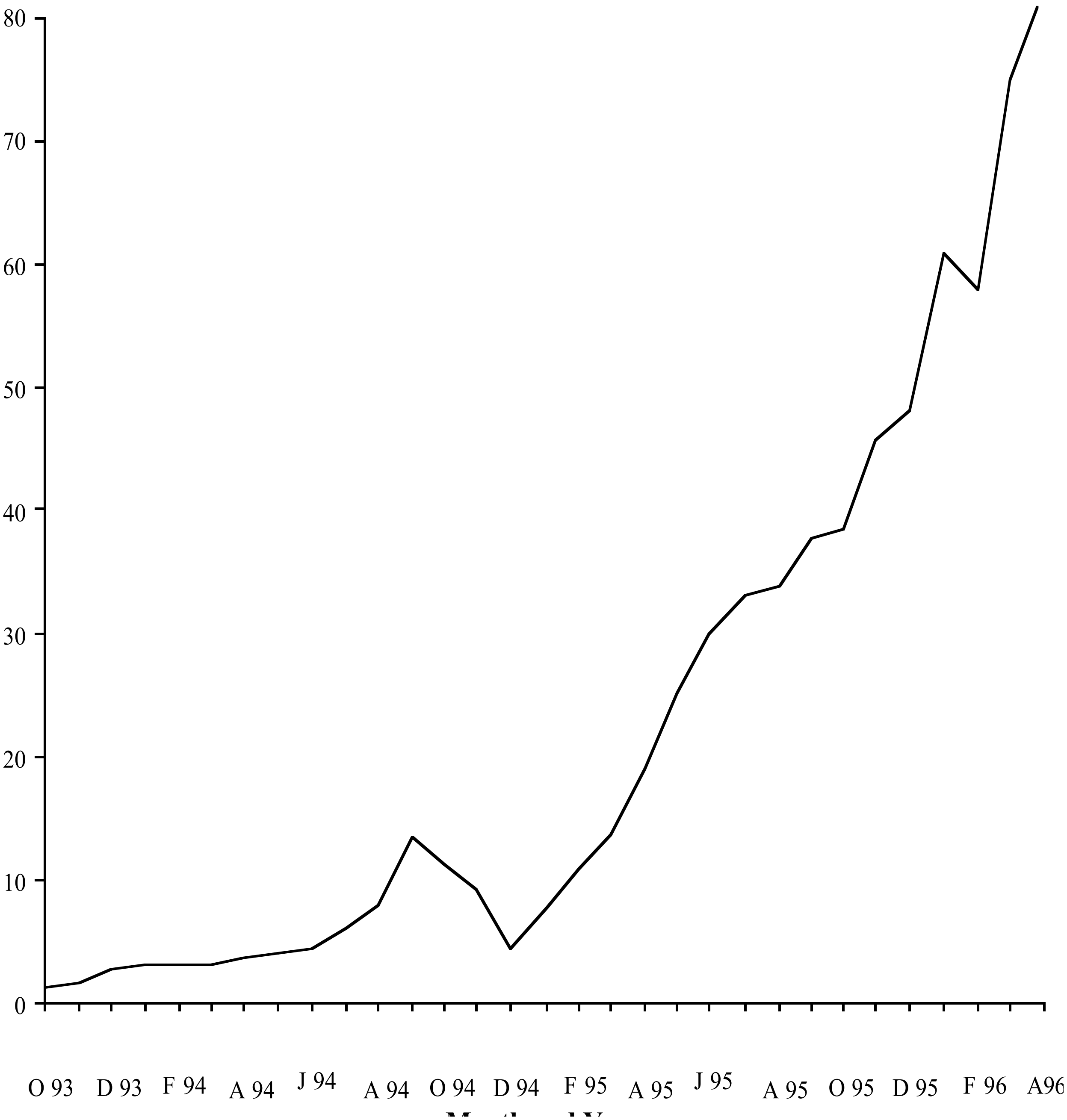
The great size of the census microdata files has always been a major obstacle to their use. Accordingly, we have developed an interactive extract system on the World Wide Web to provide easy access to the IPUMS from personal computers, workstations, and mainframe computers. The system allows researchers to fashion smaller extracts of the data specifically oriented to their own research needs and suited to their available computing power and storage capacity. In practice, researchers never require all variables and all cases from a census year. In the past, however, they have had no choice but to obtain the entire census samples to get the cases they wanted. Consequently, innumerable gigabytes of unused data are occupying tapes, hard drives, and other storage media across the country. With our extract system, researchers can design subsamples incorporating a subset of variables pertaining to the specific population(s) of interest to them.

Table 2. Descriptions and Sizes of the Public Use Microdata Samples Incorporated in the IPUMS

			Number of Records (thousands)			
Sample Description	**Year Released**	**Sample Density**	**Household**	**Person**	**Number of Variables**	**File Size**
1850 PUMS -- Free population	1994	1 in 100	37	198	92	79 Mb
1860 PUMS -- General sample*	2001	1 in 100	66	354	94	141 Mb
1870 PUMS -- General sample*	2001	1 in 100	80	428	94	170 Mb
1880 PUMS -- General sample	1994	1 in 100	107	503	123	204 Mb
1900 PUMS -- General sample	1980	1 in 760	27	100	94	43 Mb
1910 PUMS -- General sample with Hispanic and Black oversamples*	1996	varies	113	480	125	198 Mb
1920 PUMS -- General Sample	1998	1 in 100	257	1037	122	433 Mb
1940 PUMS -- General sample	1984	1 in 100	391	1351	174	584 Mb
1950 PUMS -- General sample	1984	1 in 100	461	1922	170	798 Mb
1960 PUMS -- General sample	1971	1 in 100	579	1780	141	790 Mb
1970 PUMS -- 5% State sample	1972	1 in 100	744	2030	206	929 Mb
1970 PUMS -- 15% State sample	1972	1 in 100	744	2030	210	929 Mb
1970 PUMS -- 5% County group sample	1972	1 in 100	744	2030	203	929 Mb
1970 PUMS -- 15% County group sample	1972	1 in 100	744	2030	207	929 Mb
1970 PUMS -- 5% Neighborhood sample	1972	1 in 100	744	2030	260	1016 Mb
1970 PUMS -- 15% Neighborhood sample	1972	1 in 100	744	2030	264	1016 Mb
1980 PUMS -- "A" Sample	1983	1 in 20	4711	11337	276	5376 Mb
1980 PUMS -- "B" Sample	1983	1 in 100	942	2267	276	1075 Mb
1980 PUMS -- "C" Sample	1983	1 in 100	942	2267	266	1075 Mb
1990 PUMS -- 5% Sample	1992	1 in 20	5528	12500	252	6039 Mb
1990 PUMS -- 1% Sample	1992	1 in 100	1106	2500	252	1208 Mb
1990 PUMS -- 3% Elderly sample*	1993	1 in 33	*	*	*	*
1990 PUMS -- 1% Unweighted sample	1995	1 in 100	1106	2500	252	1208 Mb
					TOTAL	25,171 Mb

* = not yet in the IPUMS. At present, the IPUMS contains the 1910 PUMS general sample and a preliminary version of

Figure 1. Volume of IPUMS Data Downloaded: Gigabytes per month, three-month moving average



Our chief goal in designing an extract interface was to simplify access to historical census microdata, a task complicated by the sheer size and complexity of the database and the differing availability of variables across samples. The challenge was designing a system that makes these complications invisible to the user.

a. Preliminary version of the interface

We have already resolved some of the most difficult design issues for developing a new generation of data extraction software. The key component is the user interface, which was developed using the programming language Perl. This is still in preliminary form, but the design incorporates many of the features we envision. Before users initiate a data extract, they are prompted for their e-mail address, which provides us with a means of contacting them and constructing a unique file name for their extract output.

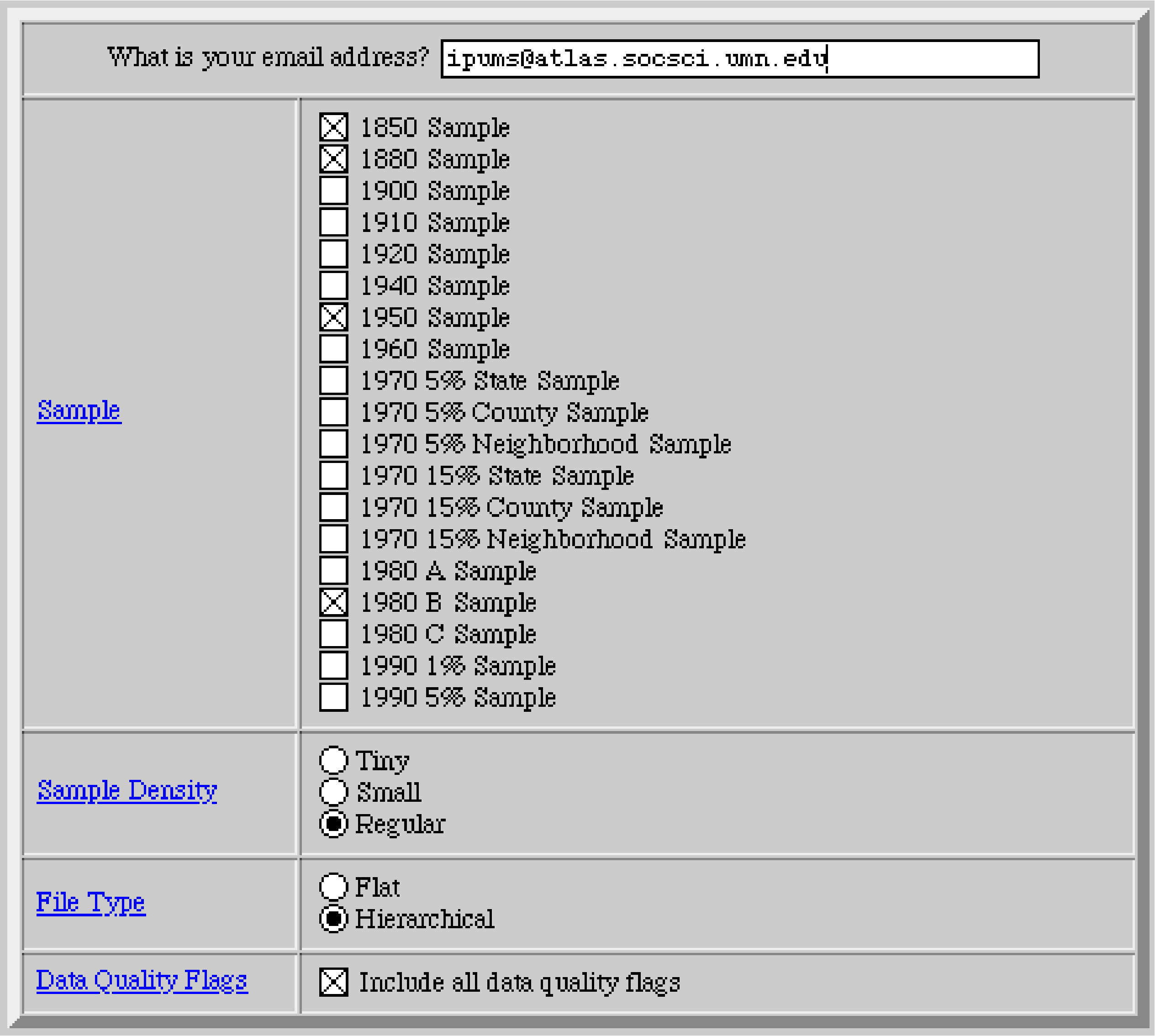
The extract procedure involves four steps—each on a separate Web page—with the contents of each page depending on selections made on the previous page. In the future, at any stage of the procedure, a query button will provide context-sensitive help explaining in detail all of the choices available to the user.

On the first page, partially shown in Figure 2, users define the general characteristics of their desired extract. They select the particular census sample or combination of samples they want (e.g., 1970 5% state sample, or the 1880 general sample) and the preferred file structure for their extract: hierarchical (household record followed by person records) or rectangular ("flat"—all household information attached to respective household members). Several sample densities are available, ranging from 1-in-20 samples available in recent census years to very small ("tiny") samples constructed in all years for purposes of testing and instruction. A feature allowing continuously variable sample densities will be added in the future. Finally, researchers may elect to include data quality flags, in which case the program will automatically append the flags corresponding to selected variables.

In the example shown in Figure 2, the user has elected to extract from the full ("regular") samples for 1850, 1880, 1950, and 1980 B. The extract will be produced in hierarchical format and will include data quality flags.

On the second page of the extract interface users select which variables they want to include in their extract. Only those variables available for the particular samples selected on the first page are displayed as options. If users have selected multiple census samples, all variables occurring in any of the specified samples are available. In all cases the form briefly describes each variable and indicates its availability among samples. Some variables have a second check box allowing users to select cases based on the value of the variable. In the future, we will add case selection boxes for many more variables. In addition, clicking on a variable name will call up all relevant documentation (see below). Users can also select entire groups of related variables by checking a single box at the end of each variable group.

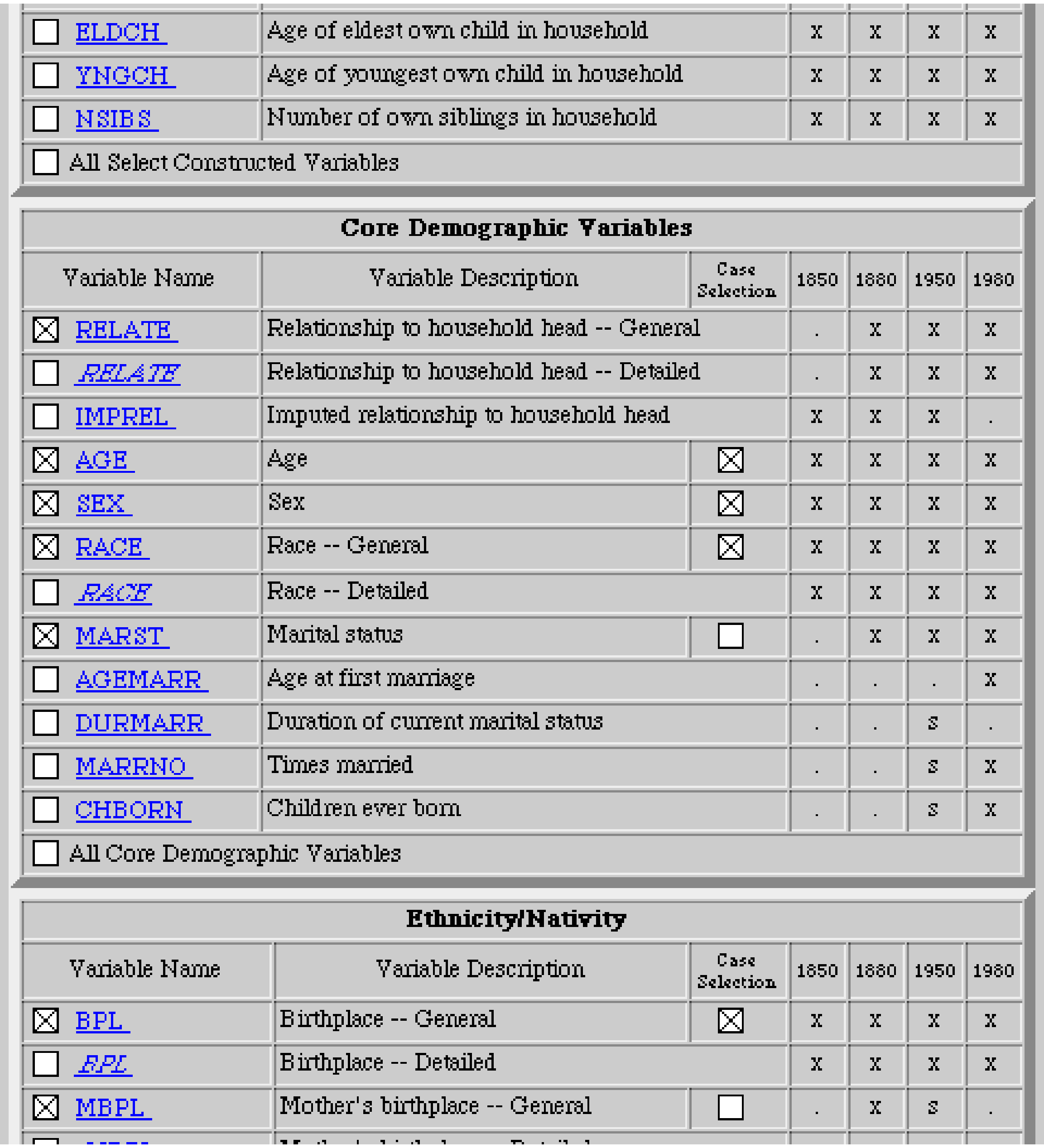
Figure 3 partially displays the selections available to a user based on the choices entered in Figure 2. Only the samples for 1850, 1880, 1950, and 1980 are shown, along with the variables available in each of those years. In this case, the user has chosen a set of basic demographic variables (checked in the left-hand column). In the case of age, sex, race, and birthplace, the user has checked the case selection box, indicating that s/he wishes to select cases based upon particular values for those variables.



What is your email address? ipums@atlas.socsci.umn.edu

Sample	☒ 1850 Sample ☒ 1880 Sample □ 1900 Sample □ 1910 Sample □ 1920 Sample □ 1940 Sample ☒ 1950 Sample □ 1960 Sample □ 1970 5% State Sample □ 1970 5% County Sample □ 1970 5% Neighborhood Sample □ 1970 15% State Sample □ 1970 15% County Sample □ 1970 15% Neighborhood Sample □ 1980 A Sample ☒ 1980 B Sample □ 1980 C Sample □ 1990 1% Sample □ 1990 5% Sample
Sample Density	○ Tiny ○ Small ● Regular
File Type	○ Flat ● Hierarchical
Data Quality Flags	☒ Include all data quality flags

Figure 2. IPUMS Sample Selection (page 1)



□ ELDCH	Age of eldest own child in household	x	x	x	x
□ YNGCH	Age of youngest own child in household	x	x	x	x
□ NSIBS	Number of own siblings in household	x	x	x	x
□ All Select Constructed Variables					

Core Demographic Variables

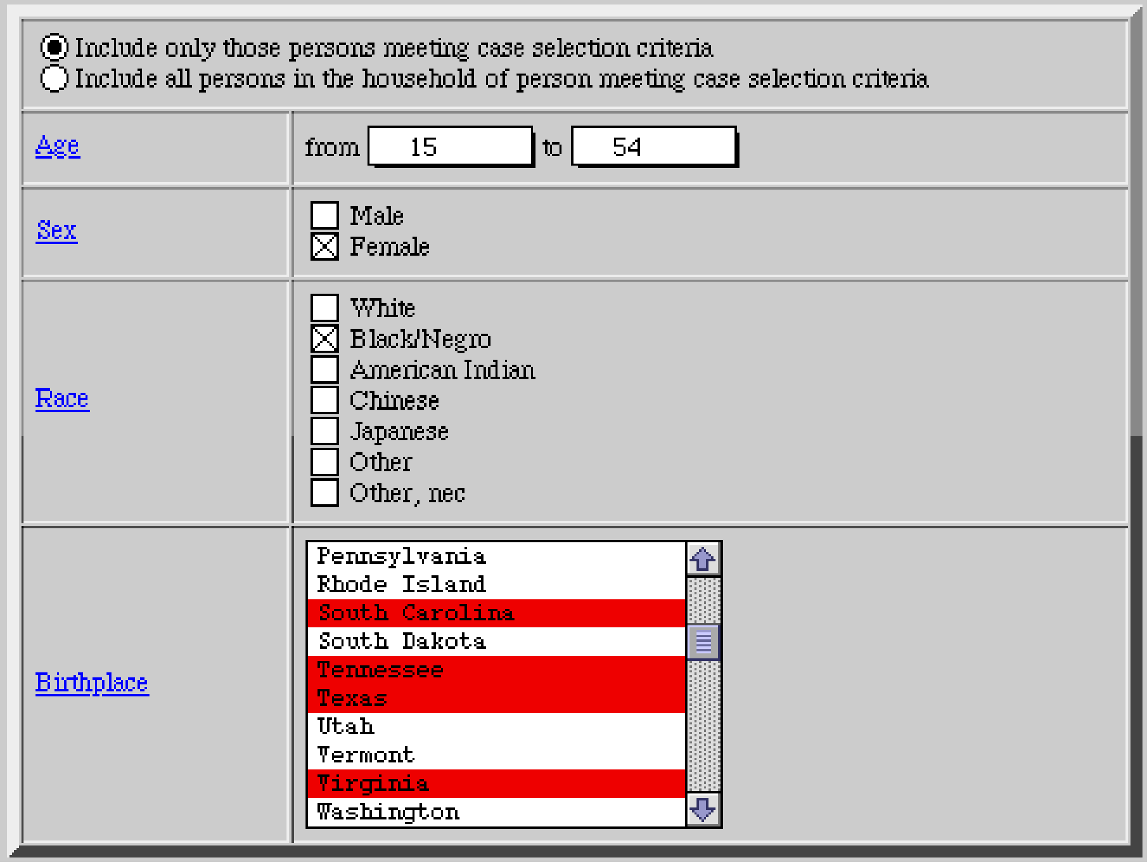
Variable Name	Variable Description	Case Selection	1850	1880	1950	1980
☒ RELATE	Relationship to household head -- General		.	x	x	x
□ *RELATE*	Relationship to household head -- Detailed		.	x	x	x
□ IMPREL	Imputed relationship to household head		x	x	x	.
☒ AGE	Age	☒	x	x	x	x
☒ SEX	Sex	☒	x	x	x	x
☒ RACE	Race -- General	☒	x	x	x	x
□ *RACE*	Race -- Detailed		x	x	x	x
☒ MARST	Marital status	□	.	x	x	x
□ AGEMARR	Age at first marriage		.	.	.	x
□ DURMARR	Duration of current marital status		.	.	s	.
□ MARRNO	Times married		.	.	s	x
□ CHBORN	Children ever born		.	.	s	x
□ All Core Demographic Variables						

Ethnicity/Nativity

Variable Name	Variable Description	Case Selection	1850	1880	1950	1980
☒ BPL	Birthplace -- General	☒	x	x	x	x
□ *BPL*	Birthplace -- Detailed		x	x	x	x
☒ MBPL	Mother's birthplace -- General	□	.	x	s	.

Figure 3. IPUMS Variable Selection (page 2, partial view)

The third page, shown in Figure 4, provides for case selection. Only those variables chosen for case selection on the second page will appear on the third. Depending on the type of variable, the page employs one of three procedures. For simple categorical variables such as region, the user selects values from a series of check boxes. With complex categorical variables such as birthplace, values are selected from a scroll-box that displays descriptive value labels rather than numeric codes. For numeric variables like age, users select minimum and maximum values. Users have the option of selecting: (1) only those individuals with the selected characteristics; or (2) entire households containing individuals with the selected characteristics.



● Include only those persons meeting case selection criteria
○ Include all persons in the household of person meeting case selection criteria

Age	from 15 to 54
Sex	□ Male ☒ Female
Race	□ White ☒ Black/Negro □ American Indian □ Chinese □ Japanese □ Other □ Other, nec
Birthplace	Pennsylvania Rhode Island 🟥 South Carolina South Dakota 🟥 Tennessee 🟥 Texas Utah Vermont 🟥 Virginia Washington

Figure 4. IPUMS Case Selection (page 3, partial view)

Extracts of the 5% samples from 1980 and 1990 are limited to cases from a single state. These files are so large that it is impractical and usually unnecessary to allow extracts on the entire samples. If either of these samples is selected, the user is forced to choose a particular state on page 3.

In Figure 4, the hypothetical user is able to select values or ranges for the variables for which the case selection box was checked in Figure 3. In this case, the researcher has chosen black women age 15 to 54 who lived in the South in the selected years. S/he has also chosen to extract only those women with the selected characteristics rather than including their entire households with them.

In the final step, users review their selections on a summary screen. If they are satisfied with their extract design, they submit it for processing. When they click the “submit” button, the program creates an extract request file that initiates the extract engine. The engine is designed to maximize input/output efficiency. Extracts are not very demanding on the processor, but are very disk-intensive.

We will inform researchers via e-mail when their extract is completed and provide instructions for downloading their files. For each extract, users receive data, codebook, and "readme" files, and an SPSS or SAS command file. The command files will contain the column locations of variables, variable labels, value labels for categorical variables, and missing values.

b. Advantages of the Minnesota approach

A powerful set of tools associated with the World Wide Web has enabled us to reduce a complex procedure to four simple steps. Using Perl allowed us to construct dynamically each page depending on the input from the previous page. This method limits choices only to valid selections, simplifying the process for users. For example, researchers who select nineteenth century census years are not offered the options for variables on the ownership of televisions or automobiles. Our preliminary user interface can be examined at:

http://www.hist.umn.edu-/~ipums/extract.html.

There have been two previous PUMS data extraction systems developed for use on the World Wide Web. One is the Census Bureau "Data Extraction System" for the 1990 1% PUMS. The other is the University of Calgary "LANDRU" system for the 3% Public Use Microdata File of the 1991 Canadian census. These efforts represent enormous strides in data dissemination, but they are incapable of handling data of the size and complexity of the IPUMS. Because of changing variable availability across years and differing sample designs, we faced a number of design issues that the Census and Calgary systems did not.

Census Bureau Web site: http://www.census.gov/ftp/pub/des/www/welcome.html

Calgary Web site: http://www.calgary.ca/~libdata/anlrud.html

Perhaps the greatest limitation of the previous extraction systems is their inability to accommodate the hierarchical structure of PUMS data. The Public Use Microdata Samples are simultaneously samples of households and of individuals, and within households the interrelationships among individuals are known. This hierarchical structure is one of the greatest strengths of the census files. By combining the characteristics of several individuals within a household, researchers can create a wide range of new variables about family and household composition and the characteristics of family members (see "advanced extract features" below). For example, we can analyze fertility by attaching the ages of all own children to their maternal records, and we can address the family economy by simultaneously measuring the age, sex, and occupation of all family members. Neither the Census Bureau system nor that developed for the Canadian census allows users to exploit directly the information contained in the structure of the data. The Census system, for example, produces extracts of either household records or person records but not both simultaneously.

We believe our extract interface is dramatically easier to use than the Census Bureau or Canadian systems. To get a basic set of demographic variables using our extract procedure requires seven selections. By comparison, a comparable extract using the Census Bureau Data Extraction System for the 1990 PUMS requires 15 to 20 times as many selections. The Bureau system is subject to certain limitations imposed by the reliance on the SAS programming language. For example, it cannot provide information about the characteristics of persons residing with a selected subpopulation. Neither the Census Bureau nor the Calgary site accommodate more than a single census year. On the other hand, both systems are somewhat more flexible than the current IPUMS system, since they allow case selection based on the value of any variable.

c. Development of a Java-based interface

Although we believe our extract interface represents a significant improvement over existing electronic extraction software, it does have limitations. In particular, it is subject to the limitations of current Web browsers (e.g., Netscape and Mosaic), which are not truly interactive. Our present extract interface is based on dynamically produced yet static forms. Each step in the process—sample, variable, and case selection—requires a separate page, the contents of each depending on selections made on previous pages. Once a page is completed, the user cannot return to it without losing later selections. For example, if the user inadvertently omits a selection of samples or variables, s/he will have to repeat all subsequent steps. Uni-directional navigation through multiple pages inevitably complicates the procedure and increases the potential for errors.

The Internet is changing rapidly, but Java is the leading candidate to become the standard protocol for transmitting dynamic and executable content over the World Wide Web in the coming years. We plan to convert the extract interface to take advantage of Java's unique capabilities. Java has several strengths that make it ideal for developing an intuitive and powerful user interface. The key advantage of Java for our purposes is its interactivity. There will be no need to navigate between static pages; a user's choices will mold a single page, which changes in real time. A truly interactive extract interface will reduce the risk of user error by simplifying the extract process. Essentially, when a user accesses our site, the extract interface program will be downloaded onto his/her computer. The client computer, not the server, runs the interface program which only accesses our site again to submit the extract request or to get additional information. This limits network traffic and demand on the server, and will make network connection speed less of a constraint.

The Java-based extract will also be suitable for CD-ROMs or other higher capacity random access storage media that may become available. The IPUMS is very large and pushes the bounds of what is practical with current transportable media. We plan to explore demand for a transportable version of the database (or a part of it) and pursue this avenue if warranted and feasible.

d. Advanced extract features

Over the last decade we have created thousands of specialized extracts from the PUMS using conventional higher-level programming languages. We realize that many users have special needs that go beyond the current capabilities of our preliminary extraction system. Accordingly, we plan to add several features to allow more complex subsample designs and the creation of new constructed variables. These include:

·A differential sample density feature that will allow researchers to select subpopulations at varying densities. For example, researchers might need to extract a subsample of 1-in-100 blacks and only 1-in-1000 whites in order to create the most efficient sample that would yield statistically significant results for both subgroups. The extract program will assign the appropriate weights to produce nationally representative statistics.

·A method for attaching characteristics of other household members to each individual's record. For example, labor economists often require information on the income and occupation of each individual's spouse. We plan to provide options for attaching any available characteristic of the household head, spouse of head, subfamily head, own spouse, own mother, and own father. The attached information (e.g., spouse's occupation) will appear on the person record as an additional variable.

·A method for counting the number of co-residing persons with any given set of characteristics. Some of these characteristics can define family interrelationships, permitting counts for groups within households such as unrelated persons, family members, or own children. Thus demographers using own-child fertility methods will be able to construct a set of variables giving the number of own children of each age for every mother. An economist could construct variables for the number of employed co-residing kin. The system will also be able to sum numeric characteristics (e.g., income or property) of select persons within households. This system, though complex, provides ample flexibility for advanced users.

2. Hypertext documentation

One of the greatest liabilities of the PUMS has always been the large initial time commitment associated with simply learning the organization of the documentation. The IPUMS mitigates but does not overcome this inherent weakness of conventional documentation. In order to address this problem, we intend to convert all of the IPUMS documentation into hypertext format using Adobe Acrobat "portable documentation format" (PDF). Along with the extract engine, we expect the hypertext documentation to revolutionize the way researchers use census data. The hypertext format will let users jump to relevant sections of the documentation with a simple click of a mouse.

The IPUMS simplifies the original PUMS codebooks, but there is no way to structure the documentation to eliminate the need to switch frequently between sections. Moreover, since the IPUMS consolidates eleven PUMS codebooks, its documentation is considerably more extensive than any one of them. In addition to the 800 page basic *User's Guide*, the documentation contains maps, comprehensive enumerator instructions, detailed descriptions of data transformations, and other elements that even modest users of the data would likely need to consult. By putting the data into hypertext with a generous number of links, users will be able to navigate the documentation with far greater ease than any previous PUMS codebook. They need never grapple with multiple volumes totaling 3000 pages or more.

The IPUMS will not only be the first census database to have hypertext documentation, it will be the first instance in which complete PUMS documentation is offered in any machine-readable form at all. Moreover, with hypertext, we can link the documentation directly to the extract interface so users can interactively make informed decisions when designing a sample appropriate for their research. By clicking on a variable name on the interface, the user will bring up the variable description and comparability discussion. Tables presenting variable frequencies suggest whether particular extracts or types of analyses are feasible in a given year. Advanced users can even look up the translation tables that detail how variables were recoded from the original PUMS into their integrated format.

The hypertext documentation will be available on our Web site and can be downloaded onto a PC, Macintosh, or UNIX system. We will also make the documentation available on CD-ROM. One of the advantages of the Adobe Acrobat PDF format is its transportability across different computing platforms. We will continue to provide heavily indexed printed and word-processor versions of the documentation that will be updated to parallel any changes made in the hypertext version.

3. User support

User support is a crucial aspect of the project. The Internet and extract system dramatically increase access to the data, but one consequence is the even more urgent need to support the growing base of users. Although we designed the extract interface to be as intuitive as possible, it will still require extensive human support. Only the most recent census years are supported by the Census Bureau. There is no institutional support for any of the earlier PUMS produced by historical researchers. Only scholars at a few major demography centers are likely to get any help at all with any but the most recent samples. Undoubtedly, the lack of sustained user support has discouraged the use of the PUMS. With the advent of the Internet and e-mail, it is now possible to centralize support for all of the census samples in their IPUMS format.

Our Web site contains a hypertext e-mail link for questions concerning the extract system, data, or documentation. In the future, the Web site will refer first-time users to an on-line tutorial that will walk them through several extract examples including variable selection, case selection, and advanced constructed variable features.

The IPUMS Web site will also serve as a repository for ancillary data such as geographic contextual or occupational wage data. For example, we have already received a number of geographic boundary files from outside researchers that translate IPUMS codes into a form suitable for mapping software. In addition, members of the census project staff have developed relevant data files in the course of their own research. A central repository for such files reduces needless duplication of effort among scholars. Researchers are invited to contribute any files they may have developed that can be attached to the IPUMS. The IPUMS already contains the necessary state and county codes to link existing county-level machine-readable statistics. Thus, for example, researchers can supplement the information in the microdata with comprehensive statistics on the racial composition and average wages for the location in which an individual resided. Such multi-level analyses are increasingly a feature of historical social scientific research (Landale and Tolnay 1991; Elman forthcoming; Ruggles forthcoming). There are also city-level machine-readable data that could be adapted to correspond with IPUMS coding.

Synopsis

Electronic communication provides a unique opportunity to disseminate the census data to a much wider spectrum of the academic community than ever before. Web navigation programs like Netscape and Mosaic, e-mail, and listservers are unprecedented resources. Despite the many files downloaded from our FTP and Web sites, however, access to the IPUMS is still largely confined to major demography centers. We anticipate that with an organized effort at dissemination, combined with our project to increase accessibility, the IPUMS data will become more widely distributed than those of any individual PUMS dataset not produced as an adjunct to the most recent census. As a result, the IPUMS can be expected to resuscitate some of the historical PUMS that have hitherto been under-utilized.

Our project, as described in this paper, is a work in progress. The completion of some aspects of the extraction system, as well as our ability to make the system publicly available and to support it adequately, depends on securing additional outside funding (the status of which is pending). We expect to proceed with development even without further funding, but the pace will be slowed and the final product may involve compromises.

The IPUMS extract system and hypertext documentation will make the census samples far more user-friendly. To this point, census microdata have generally been too cumbersome for the classroom, but our project will make the data an ideal source for instructional purposes. It will be a simple matter to custom-design extracts of the appropriate size and composition for any classroom situation.

The PUMS are a unique national resource. They are the envy of researchers from other countries, but have only been partially exploited. The IPUMS project has removed many of the barriers to using the PUMS by integrating them into a single database. But the size and complexity of the database still pose formidable obstacles to access and usability. Our project addresses both these problems, enabling researchers with very limited computing resources to take advantage of the IPUMS.

References

Elman, Cheryl (forthcoming) “Old Age, Economic Activity, and Living Arrangements in the Early 20th Century United States.” *Social Science History.*

Farley, Reynolds, and William Frey (1994) “Changes in the Segregation of Whites from Blacks.” *American Sociological Review* 59: 23-45.

Gjerde, Jon, and Anne McCants (1995) “Fertility, Marriage, and Culture: Demographic Processes Among Norwegian Immigrants to the Rural Midwest.” *Journal of Economic History* 55: 860-888.

Gordon, Linda, and Sara McLanahan (1991) “Single Parenthood in 1900.” *Journal of Family History* 16: 97-116.

Graham, Stephen N. (1980) *1900 Public Use Microdata Sample User’s Handbook.* Seattle: Center for Demography and Ecology, University of Washington.

Haines, Michael (1989) “American Fertility in Transition: New Estimates of Birth Rates in the United States, 1900-1910.” *Demography* 26: 137-148.

Hirschman, Charles, and Ellen P. Kraly (1990) “Racial and Ethnic Inequality in the United States, 1940 and 1950: The Impact of Geographic Location and Human Capital.” *International Migration Review* 24: 4-33.

Jacobs, Jerry A. (1989) “Long-Term Trends in Occupational Segregation by Sex.*”* *American Journal of Sociology* 95: 160-173.

Jenson, Leif (1991) “Secondary Earner Strategies and Family Poverty: Immigrant-Native Differentials, 1960-1980.” *International Migration Review* 25: 113-140.

Johnson, N., and S. Lean (1985) “Relative Income, Race and Fertility.” *Population Studies* 39: 99-112.

Kalmijn, Matthijs (1994) “Assortative Mating by Cultural and Economic Occupational Status.” *American Journal of Sociology* 100: 422-452.

Krivo, Lauren (1995) “Immigrant Characteristics and Hispanic-Anglo Housing Inequality.” *Demography* 32: 599-615.

Landale, Nancy S., and Stewart Tolnay (1991) “Group Differences in Economic Opportunity and the Timing of Marriage: Blacks and Whites in the Rural South 1910.” *American Sociological Review* 56: 33-45.

Magnuson, Diana (1995) “The Making of a Modern Census: The United States Population Census, 1790-1950.” Ph.D. Dissertation, University of Minnesota.

Mare, Robert D. (1991) “Five Decades of Educational Assortative Mating.” *American Sociological Review* 56: 15-32.

Morgan, S. Philip, Antonio McDaniel, Andrew T. Miller, and Samuel Preston (1993) “Racial Differences in Household Structure at the Turn of the Century.” *American Journal of Sociology* 98: 798-828.

Olson, Thomas (1991) “The Women of St. Luke’s and the Evolution of Nursing, 1892-1937.” Ph.D. Dissertation, University of Minnesota.

Ruggles, Steven (1994a) “The Origins of African-American Family Structure.” *American Sociological Review* 59: 136-151.

Ruggles, Steven (1994b) “The Transformation of American Family Structure.” *American Historical Review* 99: 103-128.

Ruggles, Steven (forthcoming) *Fragmentation of American Family Structure, 1850-1990.* Cambridge, MA: Harvard University Press.

Ruggles, Steven and Russell R. Menard (1994) “Public Use Microdata Sample of the 1880 United States Census of Population: User’s Guide and Technical Documentation. (Inter-University Consortium for Political and Social Research).

Ruggles, Steven, Russell R. Menard, Lisa Dillon, and Matt Mulcahy (1995) "1850 Public Use Microdata Sample: User's Guide." (Inter-University Consortium for Political and Social Research).

Ruggles, Steven and Matthew Sobek (1995) *Integrated Public Use Microdata Series: User's Guide* (University of Minnesota, Social History Research Laboratory).

Sanderson, Warren (1987) "Below-Replacement Fertility in Nineteenth-Century America." *Population and Development Review* 13: 305-313.

Sandefur, Gary D., and Arthur Sakamoto (1988) "American Indian Household Structure and Income." *Demography* 25: 71-80.

Sassler, Sharon (1995) "Trade-Offs in the Family: Sibling Effect on Daughters' Activities in 1910." *Demography* 32: 557-575.

Shoemaker, Nancy (1991) "The American Indian Recovery: Demography and the Family, 1900-1980." Ph.D. Dissertation, University of Minnesota.

Sorenson, Ann Marie (1989) "Husbands' and Wives' Characteristics and Fertility Decisions: A Diagonal Mobility Model." *Demography* 26: 125-135.

Strong, M. A., et al. (1989) *User's Guide Public Use Sample 1910 United States Census of Population*. Philadelphia: Population Studies Center, University of Pennsylvania.

U.S. Bureau of the Census (1972) *Public Use Microdata Samples of Basic Records from the 1970 Census: Description and Technical Documentation*. Washington, D.C.: U.S. Government Printing Office.

U.S. Bureau of the Census (1973) *Technical Documentation for the 1960 Public Use Microdata Sample.* Washington, D.C.: U.S. Government Printing Office.

U.S. Bureau of the Census (1982) *Public Use Microdata Samples of Basic Records from the 1980 Census: Description and Technical Documentation*. Washington, D.C.: U.S. Government Printing Office.

U.S. Bureau of the Census (1984a) *Census of Population, 1940: Public Use Microdata Sample Technical Documentation.* Washington, D.C.: U.S. Government Printing Office.

U.S. Bureau of the Census (1984b) *Census of Population, 1950: Public Use Microdata Sample Technical Documentation.* Washington, D.C.: U.S. Government Printing Office.

U.S. Bureau of the Census (1992) *Census of Population and Housing, 1990: Public Use Microdata Sample U.S. Documentation.* Washington, D.C.: U.S. Government Printing Office.

Watkins, Susan, ed. (1994) *After Ellis Island: Newcomers and Natives in the 1910 Census.* New York: Russell Sage.

1 Submitted for the IASSIST Conference held in Quebec, Canada. May 1995

2. Send correspondence to: IPUMS University of Minnesota Department of History, 614 SST 267 19th Ave S. Minneapolis, MN 55455 ipums@hist.umn.edu (612) 624-5818