The Minnesota Population Center Data Integration Projects: Challenges of Harmonizing Census Microdata Across Time and Place

Steven Ruggles
University of Minnesota

ABSTRACT

The Minnesota Population Center is developing three large historical census microdata series: the Integrated Public Use Microdata Series (IPUMS-USA), the International Integrated Microdata Series (IPUMS-International), and the North Atlantic Population Project (NAPP). Despite many similarities, each database presents particular challenges because of variations in source materials, organization of the projects, and institutional or legal constraints. This paper describes how the challenges we face differ across our projects.

1. Background: Origins of Public-Use Census Microdata

The first public use census microdata sample was created as a byproduct of the 1960 United States Census (U.S. Bureau of the Census 1963). In an effort to meet the needs of scholars who required specialized tabulations, the Census Bureau created a one-in-1000 extract of the basic data tapes they had used to create tabulations for the published census volumes. To preserve confidentiality, the Census Bureau removed names, addresses, and other potentially identifying information.

The 1960 public use sample revolutionized analysis of the American population and led to an explosion of new census-based research. Not only did it allow researchers to make tabulations tailored to their specific research questions; it also allowed them to apply new methods to the analysis of census data, especially multivariate techniques. But the sample did have two significant limitations. First, the sample size was relatively small. The one-in-1000 sample density yielded about 180,000 person records. Given the modest capacity of computers in 1964, this was a lot of cases, but as researchers began to use the sample for detailed analysis of small population subgroups, its limitations became apparent. Second, the 1960 public use sample provided highly limited geographic information. To ensure confidentiality, the Census Bureau stripped off all information on places below the state level. This meant, for example, that it was impossible to extract a subsample of the New York City population.

Both of these problems were addressed by the 1970 public use samples. The Census Bureau expanded the size of the 1970 samples more than 60-fold compared with the 1960 sample. The Census Bureau provided six independent public use samples for 1970, each of which had a one-in-100 density. Users who required an exceptionally large number of cases could combine the samples to obtain six-percent sample density, or about 12 million person records. In addition, the 1970 samples provided a variety of alternate geographic codes, although the Census Bureau still did not identify places of less than 250,000 population (U. S. Bureau of the Census 1972).

One additional development was critical for the long-term development of microdata in the United States. The Center for Research Libraries obtained funding from the National Science Foundation to create a new sample of the 1960 census.[1] The project was executed by DUALabs, Inc., a company headed by Jack Beresford, a former Census Bureau employee who had played a significant role in the creation of the original 1960 sample. The DUALabs version of the 1960 sample enlarged the sample density from one-in-1000 to one-in-100, and at the same time reorganized the coding schemes and record layouts to be compatible with the samples from 1970 (U.S. Census Bureau 1973). This compatibility made it relatively easy for investigators to pool data from 1960 and 1970 and thus incorporate change over time into their analyses, and this became a widespread research strategy.

By the mid-1970s, the public use samples for 1960 and 1970 had become essential tools of American social scientists. It was in this climate that Samuel Preston, then of the University of Washington, came up with the idea of creating a historical public use microdata sample by transcribing information from microfilm of census enumerator's manuscripts. Preston obtained funding from NSF in 1976 for a small (one-in-1000) sample of the 1900 census, which was completed in 1980 (Graham 1980). The original enumerator's manuscripts of the 1900 census had been publicly released in 1972, so Preston's staff had access to the original source material. When the 1910 census manuscripts were released in 1982, Preston obtained funding from both NSF and NICHD to create a somewhat larger sample of that census year (Strong et al. 1989).

Shortly after the 1900 project began, Halliman Winsborough and a group of other researchers at the University of Wisconsin further expanded the historical dimension of census microdata samples by creating machine-readable samples from the 1940 and 1950 census manuscripts. Unlike the 1900 census, the 1940 and 1950 censuses were still subject to confidentiality restrictions, so no one other than sworn Census Bureau employees was allowed to look at the original enumeration manuscripts. Therefore, with funding from NSF, Winsborough and his associates contracted with the Census Bureau to create anonymized one-percent samples from each of these census years (U.S. Bureau of the Census 1984a, 1984b).

Steven Ruggles, Rus Menard and others at the University of Minnesota picked up where Preston and Winsborough left off. In 1989, Ruggles obtained funding for a one-in-100 sample of the 1880 census manuscripts, and by the end of 1990 the Minnesota group released a preliminary 1-in-1000 version of that dataset.

---

1. NSF grant 7249358 to the Center for Research Libraries; DUALabs also received funding for the project from the Ford Foundation and NICHD (Contract 72-2707).

1.1 The IPUMS Project

By 1991 there were eight national census microdata samples available for the United States, and a ninth one—for the 1990 census—was planned (see Table 1). For the first time, the potential existed for individual-level national studies of long-run social and economic change, but in practice such analysis was cumbersome. The nine samples resulted from separate projects headed by four different investigators and six different performance sites. Even in the few cases where the same investigator working at the same location created samples for more than one census year, there were often incompatibilities, even if there was a recognizable family resemblance. For example, the two Winsborough samples (1940 and 1950) use completely different record layouts and have different coding schemes for many variables, including occupation, institution type, citizenship, and employment status. As noted, the Census Bureau samples for 1960 and 1970 were largely compatible with each other. The 1980 sample, however, was completely different, and offered substantially greater detail than did 1960 or 1970 (U.S. Census Bureau 1982).

In the late 1980s, the Social History Research Laboratory at the University of Minnesota developed a set of FORTRAN programs to extract subsets of the various historical samples and recode them into common format. This format consisted of a lowest common denominator of the level of detail available across all census years. For example, the only categories for race that were available in all the censuses were white, black, Indian, Chinese, and other, so the race variable for all census years was recoded into these five categories. The problem with the lowest-common-denominator approach is that it loses so much information that most researchers also need some variables in unrecorded format. We therefore had to prepare custom extracts tailored to the specific needs of each researcher.

These customized common-format extracts became quite popular both among Minnesota-based researchers and among a few colleagues at other institutions. By 1991 we had dedicated a small server entirely to the task of running common format extracts, and it was running continuously. We were doing this work on a volunteer basis, and we lacked resources to create documentation or verify the reliability of the recoded datasets we were producing.

In addition to harmonized codes, the IPUMS created common constructed variables. Especially useful have been the family interrelationship variables, which are "pointers" allowing researcher to link husbands with wives and parents with children without resorting to higher-level programming. We also provide compatible constructed variables describing the composition of each unit, measures of socioeconomic status, and a variety of constructed geographic variables to aid comparison across time.

The most valuable contribution of the IPUMS is the documentation. The core of the documentation is the comparability discussions, which highlight important differences and provide warnings about likely errors and strategies for enhancing compatibility for specific comparisons. For many variables, these discussions are quite long, extending up to several thousand words. In addition, we provide extensive ancillary documentation, including enumeration instructions, full detail on sample designs and sampling errors, procedural histories of each dataset, full documentation of error correction and other post-enumeration processing, and analyses of data quality.

The timing of the IPUMS project coincided with a rapidly changing technological environment. When we first conceived of the project, we expected to disseminate the data in the traditional manner up to that point: magnetic tape distributed by the Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan. By the time of our initial beta-test data release of the IPUMS in 1993, however, it had become feasible to disseminate the data via the Internet, which greatly reduced costs and improved the timeliness of data access. But the large size of the files and documentation still posed barriers for many researchers. Consequently, usage was concentrated at major universities with sufficient resources to obtain all the data and documentation and process the data.

Technological innovation presented further opportunities to democratize access to the data. In late 1995, we developed one of the earliest interactive web-based access systems for electronic dissemination of data and documentation. The user-friendly system allowed researchers to design and extract multi-year datasets containing only the variables and population subgroups needed for a particular analysis, thus greatly reducing the need for large-scale computational

Table 1. Census files incorporated in the original version of IPUMS

Census Year	Principal Investigator	Performance site
1880	Ruggles	University of Minnesota
1900	Preston	University of Washington
1910	Preston	University of Pennsylvania
1940	Winsborough	University of Wisconsin/Census Bureau
1950	Winsborough	University of Wisconsin/Census Bureau
1960	Census Bureau	DUALabs
1970	Census Bureau	Census Bureau
1980	Census Bureau	Census Bureau

resources. Essentially, the web allowed us to create an on-line version of the common-format extraction programs we had been using for years. The IPUMS dissemination system, together with the increasing power of desktop computers, brought the data within reach of virtually all academic researchers. Today, the database has almost 20,000 users, who have produced some 1,500 books, articles, dissertations, and working papers (http://ipums.org/usa/).

Since the IPUMS project began, the Census Bureau has produced new samples for 1990 and 2000 (U.S. Census Bureau 1992, 2003), and Ruggles and his colleagues have obtained funding from NIH to fill the remaining gaps in the series and to enlarge the early twentieth-century samples. When these projects are complete, the IPUMS will include samples of at least 1 percent of the American population for every census year from 1850 through 2000—with the exception of the 1890 census, which was destroyed in a fire—and will include higher-density samples at regular intervals. Table 2 describes the scope and size of the database once current and planned improvement projects are complete.

Table 2. Current and Planned IPUMS-USA Data Files

Census Year	Sample Density	Number of Records (thousands)	
		Household	Person
1850	10.0	370	1,980
1860	1.0	66	354
1870	1.0	80	428
1880	10.0	1,070	5,030
1900	5.0	1,090	4,560
1910	1.4	311	1,271
1920	1.0	257	1,037
1930	6.0	2,004	7,392
1940	1.0	391	1,351
1950	1.0	461	1,922
1960	6.0	3,474	10,680
1970	6.0	4,464	12,180
1980	9.0	8,478	20,403
1990	6.0	6,630	15,000
2000	6.0	6,792	16,884
TOTAL		35,938	100,472

1.2 IPUMS-International

Between 1960 and 1980, virtually every country in the world began using electronic computers to process census data. This ordinarily involved converting individual census responses into digital form and storing them on magnetic tape. A substantial proportion of these data still exist.

Following the release of the IPUMS in 1995 and its enthusiastic reception by the scholarly community, Robert McCaa decided that a similar effort was needed for international census microdata. A few countries—most notably Canada and the United Kingdom—had developed census microdata files modeled on the U.S. samples, but in most countries access to census microdata was restricted, expensive, or impossible.

McCaa argued that there exists a vast body of machine-readable census microdata for many countries around the world, but that it is for the most part unused. Moreover, McCaa pointed out, much of this data—especially data collected before 1985—is in danger of being lost through poor maintenance of aging nine-track tapes. In 1999, McCaa and Ruggles therefore launched the IPUMS-International initiative (Ruggles et al. 2003a). We had two principal goals: first, preservation of the World's census microdata resources, and second, democratization of access to these resources.

With major funding from the National Science Foundation (NSF SBR-9908380), we designed a demonstration project to show the feasibility of the idea. The project had four principal components:

- Inventory and Preservation. Inventory the world's surviving machine-readable census microdata and preserve them wherever possible by converting data to modern media and scanning documentation.

- Processing. Select seven countries and develop anonymized microdata files suitable for public use. This involves standardizing format and correcting format errors, drawing samples, correcting inconsistent and missing responses, assessing confidentially risks and applying protections, and harmonizing coding across countries and censuses.

- Documentation. Develop comprehensive documentation that provides guidance to users on the meaning of census responses and their comparability across time and space.

- Dissemination. Obtain licenses from national statistical agencies that allow us to disseminate microdata for educational and scholarly purposes, and distribute data and metadata through an integrated web-based data access system.

We were uncertain about the extent to which it would be possible to enlist the cooperation and support of national statistical agencies that was necessary to carry out the project. At the time we submitted our first major grant proposal, we had general letters of support from several agencies but no formal agreements and no data. Many countries were understandably cautious. Most had concerns about disclosure risk. In a few cases, statistical agencies were selling census microdata, and feared the loss of revenue if the data were made freely available.

McCaa proved to have formidable persuasive powers, and managed to convince several agency directors of the benefits of preservation and access to scientific information. Very quickly, seven countries signed up for the demonstration project. The characteristics of the census microdata from these first seven countries—chosen mainly because they signed up early, but also with an eye to geographic diversity—are shown in Table 3. We are now disseminating data from these 28

censuses via the IPUMS-International website (http://ipums.org/international).

McCaa did not stop negotiating agreements, however, and with each additional partner the job of recruiting got easier. Table 4 shows the current status of our agreements with participating countries. We have received substantial new funding to support processing of these data. In 2003 NIH funded a regionally oriented project to add most of Latin America to the IPUMS-International data series (R01 HD044154). In 2004 a second NIH grant funded a similar initiative to add a dozen countries from Europe (R01 HD047283). Most recently, IPUMS-International received a major infrastructure award from NSF (SES-0433654) which identified the project as one of the showcases of the social science division. Current funding will provide support through 2009, but additional funding will probably be required to process fully the vast data collections McCaa is acquiring.

Table 3. Current IPUMS-International Samples

Country	Census Year	Sample Density	Number of Records (thousands) Household	Person
Brazil	1960	5.0	313	3,001
	1970	5.0	1,022	4,954
	1980	5.0	1,344	5,871
	1991	5.8	2,012	8,523
	2000	6.0	2,652	10,136
China	1982	0.1	243	1,003
Colombia	1964	2.0	n.a.	350
	1973	10.0	350	1,989
	1985	10.0	571	2,643
	1993	10.0	788	3,214
France	1962	5.0	749	2,321
	1968	5.0	816	2,488
	1975	5.0	916	2,630
	1982	5.0	970	2,632
	1990	4.2	950	2,361
Kenya	1989	5.0	225	1,074
	1999	5.0	318	1,410
Mexico	1960	1.5	n.a.	505
	1970	1.0	98	483
	1990	10.0	1,648	8,118
	2000	10.6	2,312	10,099
USA	1960	1.0	579	1,800
	1970	1.0	745	2,030
	1980	5.0	4,711	11,337
	1990	5.0	5,528	12,501
	2000	5.0	6,185	14,095
Vietnam	1989	5.0	534	2,627
	1999	3.0	534	2,368
TOTAL			37,113	122,558

Table 4. Status of IPUMS-International Countries

Processing Completed			
Brazil	Kenya	Colombia	United States
China	Mexico	France	Vietnam
Data Received or Agreement Signed			
Argentina	Guatemala	Paraguay	Czech Republic
Austria	Honduras	Peru	Egypt
Belarus	Hungary	Philippines	Germany
Bolivia	Iraq	Romania	Greece
Cambodia	Ireland	South Africa	Indonesia
Chile	Malaysia	Spain	Israel
Costa Rica	Mongolia	United Kingdom	Netherlands
Dominican Republic	Nicaragua	Venezuela	Slovenia
Ecuador	Pakistan	Armenia	Tajikistan
El Salvador	Palestinian Authority	Bulgaria	Turkmenistan
Fiji	Panama	Canada	Uruguay
Under Negotiation			
Bangladesh	Iran	Nigeria	Turkey
Benin	Madagascar	Poland	Uganda
Georgia	Malawi	Russia	

1.3 North Atlantic Population Project

At an April 1999 meeting in Ottawa of the International Microdata Access Group, historical demographers from five countries made an unanticipated discovery: each of us had planned projects to create complete-count national census databases for the late nineteenth century. Researchers had been creating historical census microdata samples for some time; what was new and exciting was that these projects aimed to make data for entire national populations—not samples of populations—available to social scientists. We immediately realized that if we coordinated our activities, we could merge the datasets to create an extraordinarily powerful integrated social science database—perhaps the largest database ever created for historical social science research.

The existence of the source data for most of these countries was serendipitous. The Church of Latter-Day Saints (LDS), in collaboration with local genealogical societies, laboriously digitized three of these censuses—for Britain, Canada and the United States—to provide a resource for genealogical research. That massive project involved some 12 million hours of work by thousands of volunteers and professionals, and resulted in a verified transcription of the census information on the population of those countries in 1880 or 1881. In Iceland, nineteenth-century censuses were transcribed as part of an effort to construct genealogies for genetic research. Only in Norway was the data intended for statistical research; over the past two decades, Norwegian social scientists invested more than half a million hours in digitizing historical population data.

The result of these labors was a transcription of the characteristics of 90 million persons who resided on the North Atlantic rim in the late nineteenth century. The census in each case provides information on age, sex, marital status, family relationships, occupation, birthplace, and a range of other variables, and allows the construction of a full complement of variables describing household composition, fertility, and neighborhood and community characteristics.

In 1999, however, none of these raw data were suitable for social science research. There were literally millions of occupational titles, birthplaces, family relationships and geographic localities transcribed in four different languages. Before any of these data could be exploited by social scientists, researchers had to numerically code, classify, and document each variable. That process had already begun in Britain and Norway, but was not yet underway in the other three countries. During the course of the following year, the investigators raised funds from numerous sponsors to support the painstaking tasks of data cleaning and coding in each country.[2]

2. The funding to prepare data for the first phase of NAPP (2001-2005) was provided by the Economic and Social Research Council (UK), the Leverhulme Trust, the Essex University Research Promotion Fund, the Social Sciences and Humanities Research Council (Canada), the Harold Crabtree Foundation, the Church of Jesus Christ of Latter-Day Saints,

In June and October 2000 the investigators from each country met in Minneapolis to define the parameters of the project and develop a detailed plan of work. The participants agreed that we should not simply create compatible datasets, but rather should develop a single fully integrated database with common coding systems, constructed variables, documentation, and dissemination systems. We agreed that this ambitious plan for international collaboration would require additional funding. Although each collaborator had obtained funding to process their own national censuses, there were no resources to support the intensive collaboration that was needed to ensure that the data would be compatible across countries. We therefore proposed NAPP, which provided funding to cover costs associated with coordinating international harmonization of the data. Most of the collaboration was carried out via the Internet, but NAPP also provided funds for a series of workshops at which we hammered out solutions to the most complicated issues. The project, funded by the National Science Foundation (NSF SES-0111707), had the following goals: development of common classification systems and consistent constructed variables; documentation of comparability issues; and implementation of web-based software to provide access to the database.

The scale of our task was daunting. To give just one example, the collaborating partners coded over two million different occupational titles in four languages into a common classification scheme. To maximize cross-national consistency in coding, participants from each country coded thousands of occupational titles independently, and resolved all discrepancies in conference. Were it not for NAPP, each country would have coded occupations into a different national classification, and cross-national comparison would have been impossible.

The project has been a success. We released preliminary versions of the U.S. data in August 2003 through our web-based data access tool, the Canadian data went online in December 2003, and the British and Norwegian data in November 2004. The Icelandic data—which required substantially more work than we expected—will be incorporated into the database before the end of the project in 2005. The database is distributed through http://www.nappdata.org, and the final version—incorporating all variables described in the original proposal—will be released on schedule in July 2005 (Roberts et al. 2003a).

From the outset, the investigators regarded the integrated database of complete-count censuses as the first phase of a broader effort to develop integrated demographic data for the North Atlantic region. The second phase in the NAPP project, scheduled to begin in late 2005, has three goals: (1) expanding the chronological and geographic dimension of the database by incorporating data from additional census years for each country and adding data from Sweden; (2) coordinating national projects to link individuals between censuses, which will permit longitudinal analysis; and (3) improving NAPP variables, data editing, documentation, and web-based dissemination tools. Table 5 describes the datasets incorporated in each phase of the NAPP project. We anticipate that the second phase of the project will be complete at the end of 2010.

the University of Ottawa Research Partnerships Programme, the Norwegian Research Council, the Norwegian National Archives and the Faculty of Social Sciences of the University of Tromsø, the National Science Foundation, and the National Institute of Child Health and Human Development. In addition, Statistics Iceland and the National Archives of Iceland contributed in-kind resources.

2. Differences among the Projects

IPUMS-USA, IPUMS-International, and NAPP differ significantly with respect to source materials, administration, workflow, and legal constraints. The following sections address the implications of these differences for data integration.

2.1 Data format issues

Our experience with IPUMS-USA did not prepare us for the data format problems we have encountered with IPUMS-International and NAPP data. In general, the U.S. source data have a consistent structure: they are column-format hierarchical ASCII files consisting of a record for each household followed by a record for each person in the household. There are only a few internal inconsistencies—such as households that do not have the expected number of person records—and it does not require substantial effort to correct them.

The source data for IPUMS-International are far more challenging. Census microdata exist in a surprisingly wide range of data structures and file formats. The simplest files are rectangular, with geographic, dwelling, household, and family information replicated on each person record. More complex file structures included multiple nested record types in a single file, records stored in separate files that must be linked together, and separate files with different record layouts for various segments of the population.

The oldest datasets—those dating from the 1960s and 1970s—are often plagued by internal structural inconsistencies, a byproduct of the severe constraints on computing and data storage in those decades. Even the most recent samples, however, require substantial effort to verify that they are free of data format problems.

We begin by reformatting each sample into a hierarchical format. Any geographic or dwelling-level information is replicated on each respective household record. This data reformatting produces a standardized input structure for subsequent recoding routines. Just as important, the data manipulation often exposes problems that could not be identified from a detailed examination of data frequencies or cross-tabulations. Thus, the process of restructuring the data is an integral aspect of diagnosis and cleaning.

Experience has taught us that national statistical offices did not always verify the consistency of different hierarchical

Table 5. Phase I and Phase II NAPP datasets

Census Year	Country	Sample Density	Number of Cases (thousands) Household	Person
Existing NAPP censuses (NAPP Phase I)				
1881	Great Britain	100	6,188	29,866
1881	Canada	100	799	4,278
1870	Iceland	100	11	60
1880	Iceland	100	14	72
1901	Iceland	100	15	78
1865	Norway	100	387	1,702
1900	Norway	100	395	2,294
1880	United States	100	10,138	50,486
TOTAL EXISTING			17,933	88,764
Censuses to be added (NAPP Phase II)				
1851	Britain	2	83	398
1852	Canada	5	31	170
1871	Canada	1	13	62
1891	Canada	5	67	350
1901	Canada	5	51	265
1911	Canada	5	74	372
1921	Canada	4	74	362
1931	Canada	3	67	320
1941	Canada	3	77	355
1951	Canada	3	93	420
1703	Iceland	100	9	50
1835	Iceland	100	10	56
1845	Iceland	100	10	57
1801	Norway	100	164	879
1875	Norway*	2	135	639
1890	Sweden	100	965	4,576
1850	United States	1	37	198
1860	United States	1	66	354
1870	United States	1	80	428
1880	United States	10	1,014	5,049
1900	United States	6	1,248	5,220
1910	United States	1	311	1,271
1920	United States	1	257	1,037
1930	United States	6	1,670	6,160
TOTAL TO BE ADDED			6,632	29,120

levels within census data. We have often encountered mismatches between dwellings, households, and persons. The marginal distributions of both individual and household characteristics generally match published statistics, but inconsistencies between record types create problems for the construction of microdata samples. These include households without persons, persons without households, or households blended together. Such overt data problems rarely involve large numbers of cases; nevertheless they have to be addressed to produce clean and consistent datasets. Space constraints prevent us from describing here the full variety of data problems we encountered and explaining our solutions. Each sample is different, and we employ whatever internal data are available to arrive at a strategy for logical or probabilistic correction of errors.

The NAPP data pose different data format challenges. Most of the NAPP data were created by volunteers from the Church of

Jesus Christ of Latter-Day Saints (LDS). Over a seventeen year period (1992 to 1999) these volunteers invested eighteen million hours transcribing and retranscribing information on almost ninety million persons who resided in the United States, Canada, and Great Britain. Not surprisingly, they sometimes made mistakes. Additional errors were introduced by LDS programmers, as they attempted to merge thousands of datasets collected by thousands of different people using several generations of data-entry software. The list of errors in the LDS data is long and sordid, including omitted records, duplicated records, missing and corrupted locator keys (e.g., microfilm reel and page information, as well as variables used to order cases within a given page), misidentified household breaks, and misidentified geographic records. We have explained at length elsewhere our approach to correcting them (Goeken et al. 2003).

2.2 Harmonization of variable coding systems

The strategies and effort required for data harmonization depends in large measure on the characteristics of the source data, and these characteristics differ substantially across the three projects.

IPUMS-USA was our original model for variable-level harmonization. It was the incompatibility of variable classifications and coding systems across the U.S. samples that provided the primary impetus for the IPUMS project. We had two competing goals. On one hand, we want to keep the variables simple and easy to use for comparisons across time and space. This requires that we provide the lowest common denominator of detail that is fully comparable, with underlying complexities transparent to the user. On the other hand, we must retain all meaningful detail in each sample, even when it is unique to a single dataset.

The Census Bureau employed differing numeric classification systems in every census year, and reconciliation of these classifications was a major goal of the IPUMS. For most variables, it is impossible to construct a single uniform classification across all census years without loss of information. Since some census years provide greater detail than others, reducing all census years to the lowest common denominator would sharply reduce the power of the data series. For example, the household relationship classification for 1960 consists of only fifteen categories, compared with twenty-six categories in 1950. If we were to adopt the 1960 classification as a standard, we would lose the ability to distinguish such household relationships as nephews, aunts, and domestic employees.

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.

The data for the period before 1940 pose few problems of variable harmonization. With the replacement of the Preston samples for 1900 and 1910, the IPUMS-USA data series for the period from 1850 through 1930 is now highly consistent: all the samples were created at Minnesota using the same sample design and definitions. We have access to the original open-ended responses to most census inquiries, so codes for this early period can usually be made compatible with virtually any classification system. We have developed common data dictionaries across samples, so that we can be confident that any alphabetic string will receive the same code in every year before 1940.

The lowest common denominator for IPUMS-USA data is comparatively easy to determine. When the 1960 and 1970 samples were created in the early 1970s, data storage costs were extremely high so the length of variables was short. Indeed, the designers of the 1970 datasets were so concerned about conserving space that they packed data quality information for three different variables into each one-character data quality flag. Race was squeezed into one digit, family relationship and country of birth into two digits, and income into three digits. In most cases, therefore, the lowest common denominator for a variable is determined by the classification used in 1960 and 1970.

The samples produced at the Census Bureau since 1980—which include the 1940 and 1950 census years as well as 1980, 1990, and 2000—differ from one another in the details, but they bear a distinct family resemblance. It is comparatively easy to develop detail codes that maximize compatibility across years.

When we began to work with IPUMS-International data, we found that the model we had developed for the U.S. data did not always work. None of the samples included in IPUMS-International provide the kind of open-ended alphabetic string variables available for the U.S. before 1940; in all cases, the variables come to us coded into numeric categories. Not surprisingly, the variety of classifications systems used in the International data is far greater than is found in the samples produced by the U.S. Census Bureau. We must not only contend with the idiosyncratic practices and traditions of each statistical agency, but also with differences in language, culture, and social institutions that make interpretation of census categories difficult.

Educational attainment provides an example of a particularly difficult harmonization challenge. Our goal was to make a roughly comparable variable describing the level of schooling completed, but the source data included samples providing degrees, ones with actual number of years of schooling, and some with a mixture of both. We determined that we could consistently identify four levels of schooling completed in the first digit of "educational attainment": less than primary, primary, secondary, or tertiary. The second and third digits retain details such as differing years of primary schooling, technical versus general study tracks, and different types of degrees earned.

Several compromises were necessary to make even this rough educational scheme work across many countries. For maximum consistency, we applied the United Nations standard of six years of primary schooling, three years of lower secondary schooling, and three years of higher secondary schooling. But it was not possible to sustain these distinctions consistently across all samples because of differing national educational systems or lack of exact years completed. Moreover, some countries changed their educational systems within the time frame covered by the IPUMS samples. In the case of Kenya, which went from a 7-6-3 to an 8-4-4 system, we had to use the person's age to infer which educational system they were educated under to determine their level of schooling.

The source data for NAPP pose very different harmonization challenges. The IPUMS-International data is difficult to harmonize because the source data provide too little detail. The NAPP data is difficult to harmonize because the source data provide too *much* detail. Like the early period of IPUMS-USA, the NAPP source data are full-text transcriptions of open-ended inquiries on census enumeration forms. There is a key difference, however, between IPUMS-USA and NAPP: the IPUMS is composed of samples, and the NAPP included the entire population of five countries. Therefore, for example, instead of the 25,000 or so occupational strings we might encounter in an IPUMS sample, the NAPP contains 2,605,301. Most of these responses are in English, but NAPP also includes census responses in French, Norwegian, and Icelandic. There are not many technical obstacles to harmonizing the NAPP variables; the content of the censuses is for the most part closely comparable. The big challenges are the daunting scale of the needed classifications combined with the decentralized structure of the work process, an issue discussed in greater detail below.

2.3 Administration, work process, and legal constraints.

The three MPC microdata projects differ substantially in their organization. IPUMS-USA has the simplest structure. IPUMS-USA is located entirely in the Minnesota Population Center, and most activities are carried out in-house. Some work—such as data entry—is subcontracted to other organizations, but they do not play a significant role in the design of the database.

NAPP administration and work process represents the opposite extreme. The Minnesota Population Center coordinates NAPP, but the project is a collaboration of seven different institutions in five countries. The work—such as classifying occupations or cleaning data—is carried out separately by participants in each country. Thus, our greatest challenge is communication and coordination. Producing a single coherent database with staff and funding scattered across seven institutions on two continents requires continuous intensive communication and negotiation, and this is hard work.

Variable coding is the task that requires the closest coordination. As noted, the NAPP datasets give us access to alphabetic character strings (in English, French, Icelandic, Norwegian, or Swedish) that represent a transcription of the information collected from each individual. Each country has raised funds to classify these alphabetic strings into numerically coded categories. Some variables—age, sex and marital status—can be made comparable with little effort, but the complex variables require close collaboration to develop common coding standards. This work is difficult enough in the context of a single country; for a project of this scale, it requires a team of expert coders who work in close cooperation, sharing coding decisions continuously.

To translate from character strings into numeric codes, we construct data dictionaries that assign a numeric code to each alphabetic variation that occurs in the data. A merged dictionary containing the work of coders in each country is maintained on a central server in Minnesota. The merged NAPP dictionaries are of unprecedented scale, since they include the alphabetic strings from all six countries.

Each participant is not only responsible for coding data from their own country, but also must work to ensure consistency across countries. Subsets of the dictionaries are coded by participants in multiple countries, and any differences are worked out in conference. When possible, differences are resolved by email discussion. More controversial issues are reserved for annual meetings of the collaborators, meetings that rotate between countries.

The IPUMS-International project falls somewhere between IPUMS-USA and NAPP with respect to administration and work process. IPUMS-International has about 50 partner organizations, mostly national statistical agencies. With such a large number of organizations, a fully distributed system of work allocation on the NAPP model would be unwieldy. Accordingly, although like NAPP this is a collaboration, the work is more centralized in Minneapolis. Our partners are heavily involved in conversion of data to modern media, gathering and interpreting documentation, and sometimes translation. Data processing and harmonization tasks are usually carried out in Minnesota, but in some instances partner organizations contribute to this work.

The three projects also differ in governance. NAPP functions for the most part by consensus, but in case of disagreement all of the collaborators have agreed to abide by decisions of the majority. IPUMS-International is governed by an Oversight Board appointed by the National Science Foundation. The role of the Board to date has been largely advisory, and it is not clear whether the Investigators or the Board would prevail if there were a disagreement in policy. The relationship of the project to the partner organizations is governed by memoranda of agreement, and any disputes with national statistical agencies will be settled by arbitration under the authority of the International Court of Arbitration in Paris. The IPUMS-USA project diverges from these models, in that it is essentially a dictatorship answerable to no one.

Finally, the three projects differ with respect to the ownership of data and dissemination restrictions. The source data for IPUMS-USA are entirely in the public domain; they may be freely copied and redistributed, and there are no

confidentiality restrictions whatsoever. The Minnesota Population Center holds a copyright on the improvements and transformations we have made to the source data (including the conversion of the pre-1940 data into machine-readable form), and that allows us to impose some restrictions on users. In particular, we do not permit the data to be sold as part of a commercial product, except by permission (which has been granted to three companies). We do, however, permit the U.S. data to be used for commercial research, by journalists, or for any other purpose.

The NAPP data has multiple owners. The Norwegian, Icelandic, and British data are owned by the governments of those countries, but we have permission to distribute the data freely. The LDS has copyright on the data from the United States and Canada, and has granted us dissemination rights for non-genealogical purposes only. Accordingly, the NAPP data are restricted-access: prospective users must describe their research project, and genealogists are screened out.

IPUMS-International data are generally owned by the National Statistical Agency that created them, and we have obtained a perpetual dissemination license from each country. Access to the data is restricted to minimize risks to respondent confidentiality. Before obtaining data, individual researchers must complete an application for data access and sign an electronic license agreement (http://www.ipums.org/cgi-bin/ipumsi/ipumsireg.cgi). As part of the agreement, researchers must agree to do the following:

- Maintain the confidentiality of persons, households, and other entities. Attempts to ascertain the identity of persons or households from the microdata are prohibited, as are allegations that a person or household has been identified.

- Implement security measures to prevent unauthorized access to census microdata. Under our agreements with collaborating agencies, redistribution of the data to third parties is prohibited.

- Use the microdata exclusively for scholarly research and education. Researchers are not permitted to use the microdata for any commercial or income-generating venture.

- Report all publications based on these data to the Minnesota Population Center (MPC), which will in turn pass the information on to the relevant national statistical agencies.

In addition, researchers must propose a research project that demonstrates a scientific need for the microdata. Each application for access is evaluated by senior staff. Once an application is approved, the user password is activated, allowing controlled access to data. Penalties for violating the license include revocation of the license, recall of all microdata acquired, filing a motion of censure to the appropriate professional organizations, and civil prosecution under the relevant national or international statutes. Employees of the MPC who work with the census microdata also sign agreements to respect data confidentiality.

3. Discussion

The Minnesota Population Center is engaged in the three largest census microdata integration projects ever attempted. Superficially, the projects are very similar. Their goals are all the same: They all aim to clean, harmonize, document, and disseminate large collections of data. But because of differences in the history of the projects, differences in the nature of the source materials, and differences in their organization and administration, the projects differ dramatically.

IPUMS-USA seemed difficult when we began work fifteen years ago. Now it seems like a cakewalk. The sources are quite consistent with one another to begin with, the quality is high with few internal inconsistencies, the documentation is excellent, and the data are freely available.

IPUMS-International, by contrast, is more of an exercise in statistical archaeology. You never know exactly what you will find when you go digging into a file. The data may be scrambled or unintelligible, the documentation may be fragmentary, and it is a continuing challenge to make it all make sense.

NAPP poses fewer mysteries than IPUMS-International, but it presents great logistical challenges. The large scale of the data, consisting entirely of alphabetic strings, requires many thousands of hours of hand labor to transform into usable form. Coordination of the labor of staff in multiple countries separated by an ocean further complicates the project.

Some of these differences have implications for the data user. IPUMS-International, for example, is harder to use than IPUMS-USA or NAPP. The diversity of the source data means that IPUMS-International can never be as fully integrated as the other databases. This unavoidably places more responsibility on the user to pay close attention to the documentation, and more pressure on us to identify the major issues affecting intelligent interpretation of the data.[3]

Our next step is to bridge the differences between the projects, and provide the means for users to merge the data from NAPP,

3. We expect that IPUMS-International will eventually include about 140 samples. Given that scale, the amount of documentation for any given variable is likely to become overwhelming. To address this problem, we are designing an interface that will filter the documentation to show a subset of censuses. Thus, for example, researchers interested only in France will only see the variables and comparability discussions relevant to the French samples. We also intend to link translated and original language census questionnaires and instructions tightly to the variables to encourage users to examine the original wording and layout of the questions.

IPUMS-USA, and IPUMS-International. Cross-project comparisons will follow two strategies:

IPUMS-USA Format. Most NAPP variables are already coded to the IPUMS-USA standard, with the notable exception of occupation. We are developing a compatible occupational variable by coding all US data into the standard NAPP classification (Roberts et al. 2003b). When this is complete, it will be feasible to combine data from NAPP and IPUMS-USA, and we plan to make merged data extracts available from either website.

IPUMS-International Format. We have already converted IPUMS-USA samples for the period since 1960 into IPUMS-International format, and it will be a comparatively simple task to recode the earlier IPUMS-USA samples and the NAPP database into the IPUMS-International system. We will then allow users to merge consistently-coded data from all three projects.

The merged data will never be seamless. We are dealing with hundreds of censuses taken in some 50 countries over the course of two centuries, and we cannot make the real differences disappear. But we can reduce some of the barriers to cross-national and cross-temporal research. Each project is preserving datasets and making them freely available, converting them into a uniform format, providing comprehensive documentation, and implementing web-based tools for disseminating the microdata and documentation. If we do our job well, we will stimulate research that makes broad comparisons across time and space.

References

Goeken, Ron, Cuong Nguyen, Steven Ruggles, and Walter Sargent. 2003. "The 1880 United States Population Database." Historical Methods 36:1, 27-34.

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

Roberts, Evan, Steven Ruggles, Lisa Dillon, Ólöf Gardarsdottir, Jan Oldervoll, Gunnar Thorvaldsen and Matthew Woollard. 2003a. "The North Atlantic Population Project: An Overview," Historical Methods, 36:2, 80-88

Roberts, Evan, Matthew Woollard, Chad Ronnander, Lisa Dillon, and Gunnar Thorvaldsen. 2003b. "Occupational Classification in the the North Atlantic Population Project," Historical Methods, 36:2, 89-96.

Ruggles, Steven, Miriam King, Deborah Levison, Robert McCaa, and Matthew Sobek. 2003a. "IPUMS-International" Historical Methods, 36:2, 60-65.

Ruggles, Steven, Matthew Sobek, Miriam L. King, Carolyn Liebler, and Catherine Fitch. 2003b. "IPUMS Redesign" Historical Methods 36:1, 9-21.

Ruggles, Steven, Matthew Sobek, Trent Alexander, Catherine A. Fitch, Ronald Goeken, Patricia Kelly Hall, Miriam King, and Chad Ronnander. 2004. Integrated Public Use Microdata Series: Version 3.0. Minneapolis, MN: Minnesota Population Center.

Strong, Michael 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. Census Bureau. 1963. Census of population and housing, 1960 public use sample: one-in-one-thousand sample. Washington, D.C.: U.S. Government Printing Office.

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

———. 1973. Technical Documentation for the 1960 Public Use Microdata Sample. Washington, D.C.: U.S. Government Printing Office.

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

———. 1984a. Census of Population, 1940: Public Use Microdata Sample Technical Documentation. Washington, D.C.: U.S. Government Printing Office.

———. 1984b. Census of Population, 1950: Public Use Microdata Sample Technical Documentation. Washington, D.C.: U.S. Government Printing Office.

_____. 1992. Census of Population and Housing, 1990: Public Use Microdata Sample U.S. Documentation. Washington, D.C.: U.S. Government Printing Office.

_____. 2003. Census 2000, Public Use Microdata Sample (PUMS), United States, Technical Documentation. Accessed at http://www.census.gov/prod/cen2000/doc/pums.pdf, October 25, 2003.