THE IPUMS-INTERNATIONAL PROJECT: CHALLENGES AND METHODS OF INTERNATIONAL CENSUS DATA INTEGRATION

Matthew Sobek
Albert Esteve
Robert McCaa

THE IPUMS-INTERNATIONAL PROJECT: CHALLENGES AND METHODS OF INTERNATIONAL CENSUS DATA INTEGRATION

Matthew Sobek
Albert Esteve
Robert McCaa

239

Aquesta comunicació es va presentar a la Social Science History Association Conference,
St. Louis (USA) Octubre 24-27 2002

Centre d'Estudis Demogràfics

2004

INDEX

TABLE INDEX

FIGURES INDEX

THE IPUMS-INTERNATIONAL PROJECT: CHALLENGES AND METHODS OF INTERNATIONAL CENSUS DATA INTEGRATION.

Abstrat.

IPUMS-International is a project to inventory, preserve, harmonize and disseminate census microdata from around the world. IPUMS-International makes cross-national census microdata readily accessible and usable. The project facilitates comparative international research based on pooled microdata. A harmonized composite coding system for variables allows comparisons over time and across countries without loss of information. A series of constructed pointer variables allows users to identify relationships among spouses, parents, and children. Extensive documentaton aids in the interpretation of data and highlights major compatibility issues A web-based dissemination system provides convenient and free access to both the microdata and the documentation.

Key words: IPUMS, microdata, inventory, census.

Resum.

IPUMS-International és un projecte per inventariar, preservar, homogeneitzar i difondre microdates de censos de tot el món. IPUMS-International fa que les dades dels censos nacionals siguin accessible i utilitzades arreu. D'aquesta manera el projecte, facilita comparacions internacionals de les dades censals. Mitjançant la creació d'un sistema de codis per tal d'homogeneitzar totes les variables, ens permet comparar diferents països i diferents periodes de temps sense perdre, en cap moment, informació. D'aquesta manera es poden construir series per tal d'identificar les relacions de parella, familiars, fills,… S'adjunta, també, una important documentació per tal d'interpretar correctament les dades ofertes. El sistema de difusió de dades, és mitjançant una web de lliure accès a les dades i la documentació.

Paraules clau: IPUMS, microdata, inventari, censos.

Resumen.

IPUMS-International es un proyecto para inventariar, preservar, homogeneizar y transmitir los microdatos de los censos de todo el mundo. IPUMS-International hace que los datos de los censos nacionales sean accesibles a todos. De este modo, el proyecto, proporciona comparaciones internacionales de los datos censales. Mediante la creación de un sistema de códigos para la homogeneización de todas las variables, nos permite comparar distintos países y distintos periodos de tiempo sin perder, en ningún momento, información. Así, se pueden construir series para la identificación de relaciones de pareja, familiares, hijos,… También se adjunta una importante documentación para la correcta interpretación de los datos. El sistema para la difusión de los datos, es mediante una web de libre acceso a los datos y a la documentación.

Palabras clave: IPUMS, microdata, inventario, censos.

Census microdata are an invaluable resource for social science research. By providing the individual responses of discrete persons and households, microdata are immensely flexible, and can often be used for purposes not envisioned by the statistical offices that carried out the original censuses. No alternate data source on demographic and economic behavior offers comparable sample size, chronological depth, and geographic coverage. For much of the world, however, census microdata are either unavailable or highly restricted, and are therefore seldom used.

The IPUMS-International project (Integrated Public Use Microdata Series – International) addresses this shortcoming by providing access to large census samples from five continents. The project aims not simply to make international census data available, but to make them usable. Even where census microdata can currently be obtained, comparison across countries or time periods is challenging because of inconsistencies between datasets and inadequate documentation of comparability issues. Because of this, comparative international research based on pooled census samples is rarely attempted, even in cases where the research topic clearly calls for such an approach. IPUMS-International reduces the barriers to international research by converting world census microdata into a uniform format, providing comprehensive documentation, and by making the data available to researchers through a web-based access system.

In the United States, census microdata samples have been available to researchers for almost forty years and have become an indispensable component of social science infrastructure. The retroactive creation of new, older samples for 1850 to 1950 has further increased the richness of the U.S. data and encouraged more historically oriented research. The Integrated Public Use Microdata Series (IPUMS-USA) is partly responsible for the widespread use of census microdata by researchers studying the United States. IPUMS-USA makes census microdata from 1850 to 1990 freely available to scholars in harmonized format through a user-friendly data access system (Ruggles and Sobek 1997; http://ipums.org/usa). Since its preliminary release in 1995, the IPUMS has become one of the most widely used demographic resources in the world, with over 6,000 registered users producing over 250 articles and books to date.

The IPUMS-International project extends the IPUMS model beyond the United States. Begun in 1999, IPUMS-International takes microdata samples from eight countries with

broad geographical distribution and cleans, harmonizes, documents, and disseminates them using the same principles and methods that underlie the original IPUMS-USA database. The five-year project is funded by the National Science Foundation, supplemented by a grant from the National Institutes of Health[1].

State of the Project

IPUMS-International is now completing its third year of work. In the first year of the project we carried out a comprehensive inventory of known microdata, much of which is described in the *Handbook of International Historical Microdata* (Hall, McCaa, and Thorvaldsen 2000). At the same time, the United Nations Statistics Division's historical archive of census documentation was transferred to the Minnesota Population Center for archiving. Census enumeration forms in the collection were scanned, totaling over 400 censuses from the 1950s to the present. The first release is available on CD and from the IPUMS-International web site[2]. In May 2002, we released the first preliminary group of harmonized census microdata samples for Colombia, France, Kenya, Mexico, the United States, and Vietnam. The samples are incorporated within a betatest web-based data access and documentation system (http://ipums.org/international). Only a subset of the available variables are included. The online system has remained largely unchanged since May, while we continue developing the database behind the scenes in preparation for improvements and expansion in 2003 and 2004.

The current extract system, although rudimentary, has the basic functions we envision. Due to confidentiality restrictions, researchers must apply to become registered to use the system. Once registered, users can create data extracts that contain only the samples and variables of interest to them. The extraction system is described in greater detail below. The remainder of the web site provides information on the samples and variables. Of particular note are the variable comparability discussions. These are designed to indicate where there are notable issues for interpreting a variable's codes for purposes of temporal and spatial comparison. In addition to these discussions, the web site contains the original census

[1] IPUMS-International proper is funded by a grant from the National Institutes of Health, Steven Ruggles, Principal Investigator (R01 HD037508). The procurement and development of Colombian samples is funded by the National Science Foundation, Robert McCaa, Principal Investigator (SBR 9907416).
[2] See http://www.ipums.org/international/enumeration_forms.shtml

questionnaires and instructions so users can examine the full text from the original enumerations.

Table 1 shows the samples that are currently projected to be included in IPUMS-International within its five-year grant period. The countries were chosen for their data availability at the outset of the project and their dispersed geographic coverage. In embracing such diversity one of our goals was to confront in as few samples as possible most of the key variations we would eventually encounter around the world. A second criterion was to select countries that had two or more samples in order to encourage research with chronological depth. China is included because of its inherent importance and because of the prospect for additional samples in the future.

The first release of international census microdata samples has been publicized basically by word-of-mouth. Nevertheless, we have already received over 150 applications for access to the data from scholars around the world, including representatives of four national statistical offices, the World Bank, and the World Health Organization. The topics proposed include analysis of the living arrangements of the aged, female labor-force participation and educational attainment, regional inequality differentials, the brain drain, the demographic and spatial dimensions of violence in Colombia, the relationship of disease factors to education, migration between Mexico and the United States, and the relationship of marriage to education, and the evolution of marriage patterns in Mexico.

In the final two years of the project we will add all remaining variables contained in the various IPUMS-International datasets. These datasets will include 1982 China and 5 Brazilian samples from 1960 to 2001, and depending on time and resources, possibly Hungary, Ghana, Palestine, or Spain. Aside from U.S. Census 2000, we will add samples to increase the number of U.S. cases for 1970 to 1990. The addition of pre-1960 U.S. samples is also a possibility, depending on time and perceived demand.

Data Preservation

In addition to identifying world microdata holdings and integrating selected countries, IPUMS-International is devoted to data preservation. The project has funded the preservation of microdata from over one hundred censuses. Many of these are samples that we are not authorized to distribute at this time, but which were in danger of soon becoming

Table 1. IPUMS-International Samples

Country	Census Year	% Sample	Persons (000s)	Households (000s)
2002 Data Release				
Colombia	1964	2	350	n.a.
	1973	10	1,989	350
	1985	10	2,643	571
	1993	10	3,274	788
France	1962	5	2,321	n.a.
	1968	5	2,488	n.a.
	1975	5	2,629	n.a.
	1982	5	2,714	n.a.
	1990	4.2	2,361	n.a.
Kenya	1989	5	1,074	225
	1999	5	1,410	318
Mexico	1960	1.5	503	n.a.
	1970	1	483	98
	1990	1	803	164
	2000	10.6	10,099	2,312
United States	1960	1	1,800	579
	1970	1 *	2,030	744
	1980	1 *	2,267	942
	1990	1 *	2,500	1,106
Vietnam	1989	5	2,627	534
	1999	3	2,368	534
2004 Data Release				
Brazil	1960	1 *	914	n.a.
	1970	1 *	1,105	n.a.
	1980	3 *	3,526	n.a.
	1991	2.7 *	n.a.	n.a.
	2001	n.a.	n.a.	n.a.
China	1982	.1	1,002	242
United States	2000	1 and 5	16,884	6,954

Notes: Possible additional samples include: Ghana 1984, 2000; Hungary 1980, 1990, 2001; Palestine 1997; Spain 1981, 1991, 2001. The French datasets currently do not include households, but that will be rectified in 2003.

* By March 2004, sample densities will be increased to 6% for 1970-1990 United States. Sample densities for Brazil may increase.

irrevocably lost. Much of this preservation effort has been carried out via the UN Demographic Center for Latin America and the Caribbean (CELADE), which has vast historical data holdings for Latin America. IPUMS-International funded the migration by CELADE of dozens of datasets and their documentation to modern media. The IPUMS-International project also maintains an off-line secure archive of data not intended for distribution, with the sole purpose of ensuring the data's survival in the long run.

Future of IPUMS-International

The current version of IPUMS-International represents the beginning of a much larger enterprise to provide access to the world's microdata. With future additional funding we hope to expand the geographic coverage greatly as we include more countries. At the same time, we are keen to exploit any and all opportunities to add historical samples for periods before the mid-20th century.

We have submitted a proposal to greatly expand IPUMS-International by integrating 12 countries from Latin America. With the help of CELADE, we already have access to the necessary data and have distribution agreements with 16 of 18 Latin American countries. For Argentina, there are even census samples for as early as 1869 and 1895.

We have also been scouring the rest of the world for opportunities to distribute microdata samples. Those openings are proving much greater than we supposed when we first proposed the IPUMS-International project. Among the possibilities we are pursuing is the prospect of more recent and larger samples of China. In 2003 we are likely to seek funding for a further major addition: IPUMS-Europe. We have received interest in participation from all parts of the world. Some countries have even donated their data to the project in the hope that we will at some point acquire funding to integrate it. The most surprising aspect of our data inquiries has been the number of positive reactions from countries with reputations for restrictive or non-existent access to microdata.

As the database grows in the future, we expect to add regional modules, the first of which will be for Latin America. In some cases, incompatibilities across continents are so great that variable coding schemes designed to incorporate all variations will be significantly more cumbersome than the original variable coding design. The regional classifications will take advantage of commonality in social structure and similarity in census questionnaires within regions to create more streamlined classifications. For example, we might create a

marital status variable specific to Latin America that will emphasize consensual unions and ignore such categories as polygamous marriages. The world-compatible variables will be presented side-by-side with the regional variants so researchers can choose the optimal version for their purposes.

Challenges of Data Access, Quality, and Harmonization

The remainder of this paper describes the major elements of the project: the challenges, methods and lessons learned in applying the IPUMS model to international census integration. It should be noted that the text sometimes speaks to facets of the project that have not yet been implemented, but which are part of the five-year plan.

Data Access

Privacy considerations are a growing international concern. In contrast to the United States, where the census public use samples are available without any restrictions, the international samples require satisfying the confidentiality requirements of the various national statistical agencies. To meet these needs we use two strategies for safeguarding the confidentiality of the microdata: confidentiality agreements and statistical disclosure protections.

We disseminate microdata only under strict confidentiality controls approved by each national statistical office. Before data are released, individual researchers must complete an application for data access and sign an electronic license agreement. As part of the agreement, researchers must agree to a number of conditions intended to ensure confidentiality and non-commercial use of the data. In addition, researchers must propose a research project that demonstrates a scientific need for the microdata. Once an application is approved, the user password is activated, allowing controlled access to data.

Technical safeguards of anonymization supplement administrative methods of maintaining statistical confidentiality. We work with each country's statistical office to minimize the risk of disclosing respondent information. The details of the confidentiality protections vary across countries, but in all cases, names and detailed geographic information are suppressed. In addition, a variety of other procedures are used to enhance confidentiality protection, including the following:

- Swapping an undisclosed fraction of records from one administrative district to another to make positive identification of individuals impossible.
- Randomizing the sequence of households within districts to disguise the order in which individuals were enumerated.
- Combining codes that reveal sensitive characteristics or identify very small population subgroups (e.g., grouping together small ethnic categories).
- Top coding, bottom coding, and rounding continuous variables to prevent identification.

In addition to these basic measures, we are continuing to evaluate emerging methods and technologies for disclosure protection (McCaa and Ruggles 2002, Ruggles 2000). The safety record for public use census microdata is apparently perfect. In almost four decades of use, there has not been a single verified breach of confidentiality. The IPUMS-International procedures are designed to extend this record.

Dataset Reformatting

We carry out a systematic program of data reformatting and cleaning for each dataset. This includes analysis of the record structure, reformatting of the data into a standard hierarchical format, internal consistency checks, and correction of data errors.

The international census samples exist in a variety of data formats and have various irregularities. Cleaning the data to make them suitable for public-use microdata samples proved considerably more time-intensive than we had anticipated given our past work on the U.S. datasets. The oldest international datasets we have treated to date—those dating from the 1960s and 1970s—generally pose the greatest problems, a consequence of the computing and data storage constraints of those decades. Even the most recent samples, however, require a substantial investment to verify that they are free of data format issues. In the seventeen international censuses we have processed to date, data format problems affected only a tiny fraction of cases; nevertheless, these had to be addressed systematically to produce clean sample data.

The raw data files are preserved in a remarkable variety of formats. *Rectangular* files are the simplest format, with geographic, dwelling, household, and family information replicated on each person record. In *hierarchical* files, the microdata have as many as four

nested record types identifying the starting points of each geographic area, dwelling, and household. In these files, any irregularity in the sequence of record types can create widespread data problems. *Linked* censuses are organized into multiple record types stored in separate files designed to be linked together by means of a common identification (ID) number. These record types can include mortality, fertility, and group quarters records as well as person, household, and dwelling records. Small imperfections in the ID numbers can cause significant problems. We begin by reformatting each sample into a simple, consistent hierarchical format consisting of a household record followed by person records for each individual in the household. Any geographic or dwelling-level information is replicated on each household record. This reformatting often exposes problems that cannot 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.

We cannot describe here the wide variety of data format problems we have encountered and explain our solutions. Each census is different, and we employ whatever internal data are available to arrive at a strategy for logical or probabilistic correction of errors. One of the more promising methods, however, pertains to the Colombian census of 1973. We began with the 100 percent population microdata used to create published tabulations. The data were in separate household and person files, and attempts at matching the files by household identification (ID) number uncovered an array of data errors affecting almost 3% of the records. After marking the cases with data problems, we drew a 10% sample by substituting nearby untainted records when a sample point fell on a bad record. By identifying donor households in close geographic proximity to the corrupted households, we were able to maintain representativeness. There are no detectable systematic biases in the completed 10 percent sample.

We expect to use this procedure in future cases where there are significant data integrity issues and we have complete-count or high-density data from which to draw a sample of lesser density. Such situations are more likely to arise than we imagined when we began this project. In many cases old full-count data survive, but the country in question never created a sample and now lacks the resources to do so. In Latin America, we have found widespread willingness to allow us to draw public-use samples from these full-count data, and we expect further such opportunities to arise elsewhere.

Data Cleaning

We have developed a battery of tests to ensure data soundness. While most of the datasets are generally of high quality, unlike modern U.S. census data, many have never been cleaned. Among the things we check for are households with no heads or multiple heads; households with multiple wives in countries that do not practice polygamy; implausibly large households or dwellings; and duplicate records. We also look for inconsistencies between household and person records, in the relationships among the persons in a household, and among the characteristics of individuals. For example, we check for contradictions between age and labor force status, marital status, educational attainment, and school attendance. Where data errors can be unambiguously identified, we flag the data item as inconsistent.

Once the consistency checks are completed, we edit missing and inconsistent values. Missing and inconsistent values are routinely replaced with allocated values in recent U.S. census data by means of logical edits and probabilistic hot deck imputation procedures. For example, if sex is missing, it is edited by logical inference from the family relationship field or based on the sex of a spouse.

When missing or inconsistent items cannot be resolved through logical computer editing, we turn to probabilistic allocation procedures modeled on those of the U.S. Census Bureau. Allocation of missing and inconsistent data significantly increases the reliability of sample estimates and makes the samples simpler to use. Missing data allocation is not, however, routinely incorporated in non-U.S. microdata. In allocation, for each variable there is a series of criteria for matching a donor record used to impute the missing or inconsistent value. These criteria are determined through analysis of the best predictors for each variable, and can vary from census to census and between countries. For example, if school attendance is missing, then one might allocate the school attendance of the most proximate individual in the file who shares the same age, sex, and parental occupation or income. A data quality flag identifies any allocated or edited data items.

Variable Harmonization

The first step in the integration process is to determine the availability of variables across all samples. With hundreds of variables in numerous original languages, this is not a trivial

task. Moreover, language differences aside, equating seemingly similar variables is not always straightforward because of varying conceptual and terminological conventions. Despite United Nations and other international influences, even comparable variables may be referred to differently across countries. For example, the "class of worker" variable in the United States (i.e., employer, self-employed, employee) is referred to as "status in employment" in countries that subscribe to the UN census terminology. "Status in employment," in turn, reads like the conceptually different U.S. variable "employment status" (labor force status and unemployment), which is referred to in UN terminology as "activity status".

Table 2 presents a partial listing of available variables among the six countries in the first IPUMS-International data release. (Many are not yet incorporated in the database accessible through our web site.) The variables are organized according to the UN categorization of topics, and a particular "variable" in the table may actually represent multiple variables on that topic. Variable availability differs by country and year but, in general, national differences are more important than differences within countries over time.

The international census samples employ differing numeric classification systems, and reconciliation of these codes is a major part of the work of IPUMS-International. Variable design often influences the analytical strategies adopted by researchers, and therefore must be developed with care. The aim is to create a comparable set of codes for each variable that mean the same thing across countries and over time.

The IPUMS-International design strategy is ambitious. We retain all the detail provided in the original samples. At the same time we provide a truly integrated database, in which identical categories in different census samples always receive identical codes. We employ several approaches to achieve these competing goals. In some cases, the original variables are compatible and recoding them into a common classification is straightforward. In this situation, the documentation notes any subtle distinctions between censuses. For most variables, however, it is impossible to construct a single uniform classification without losing information. Some samples provide far more detail than others, so the lowest common denominator of all samples inevitably loses important information. In these cases, we construct composite coding schemes. The first one or two digits of the code provide

Table 2. Selected Variable Topic Availability, by Country and Census Year

	Colombia				France					Kenya		Mexico				United States				Vietnam	
	64	73	85	93	62	68	75	82	90	89	99	60	70	90	00	60	70	80	90	89	99
Geography and internal migration																					
Place of usual residence	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Place of birth	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	.	.
Duration of residence	x	x	.	.	.	.	.	.	.	.	x	x	x	.	.	x	x	x	x	.	.
Place of previous residence	x	x	.	.	.	.	.	.	.	.	.	x	x	.	.	.	.	.	.	.	.
Place of residence at a specified date in the past	.	.	x	x	x	x	x	x	x	x	x	.	.	x	x	x	x	x	x	x	x
Household and family structure																					
Relationship to head of household/householder	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Demographic and social																					
Sex	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Age	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Marital Status	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Citizenship	.	.	.	.	x	x	x	x	x	x	x	x	.	.	.	.	x	x	x	.	.
Religion	.	.	.	.	.	.	.	.	.	.	x	x	x	x	x	.	.	.	.	.	x
Language	.	.	.	.	.	.	.	.	.	.	.	.	x	x	x	.	.	x	x	.	.
National and/or ethnic group	.	.	.	x	.	.	.	.	.	x	x	x	.	.	x	x	x	x	x	x	x
Fertility and mortality																					
Children ever born	.	x	x	x	.	.	.	.	.	x	x	x	x	x	x	x	x	x	x	x	x
Children living	.	x	x	x	.	.	.	.	.	x	x	.	.	.	x	.	.	.	.	x	x
Date of birth of last child born alive	.	x	.	x	.	.	.	.	.	x	x	.	.	.	x	.	.	.	.	x	x
Deaths in the past 12 months	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x
Maternal or paternal orphanhood	.	.	.	.	.	.	.	.	.	x	x	.	.	.	.	.	.	.	.	.	.
Age, date or duration of first marriage	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	x	x	.	.	.
Education																					
Literacy	x	x	x	x	.	.	.	.	.	x	.	x	x	x	x	.	.	.	.	x	x
School attendance	.	x	x	x	.	.	.	.	.	x	x	.	x	x	x	x	x	x	x	x	x
Educational attainment	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Field of education and educational qualification	.	.	.	.	x	x	.	.	.	.	.	.	.	.	x	.	.	.	.	x	x
Economics																					
Activity status	x	x	x	x	x	x	x	x	x	x	x	.	x	x	x	x	x	x	x	x	x
Time worked	x	.	x	.	.	.	.	.	.	.	.	x	x	x	x	x	x	x	x	.	.
Occupation	x	x	.		x	x	x	x	x	x	.	x	x	x	x	x	x	x	x	x	x
Industry	x	x	.	x	x	x	x	x	x	.	.	x	x	x	x	x	x	x	x	x	x
Status in employment	x	x	x	x	x	x	x	x	x	x	.	x	x	x	x	x	x	x	x	.	.
Income	.	.	.	.	.	.	.	.	.	.	.	.	x	x	x	x	x	x	x	.	.
Institutional sector of employment	.	.	.	.	x	x	x	x	x	.	.	.	.	.	.	.	.	.	.	.	x
Place of work	.	.	.	.	x	x	x	x	x	.	.	.	.	.	x	x	x	x	x	.	.
International migration																					
Country of birth	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	.	.
Citizenship	.	.	.	.	x	x	x	x	x	.	.	.	.	.	.	x	x	x	x	.	.
Year or period of arrival	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	x	x	x	.	.
Disability																					
Disability	.	.	.	x	x	.	.	.	.	.	.	.	.	.	x	.	x	x	x	.	.
Cause of disability	.	.	.	.	x	.	.	.	.	.	.	.	.	.	x	.	.	.	.	.	.

Notes: Samples are identified by the last two-digits of their census year. An "x" indicates the topic is available in that sample.

information available across all samples. The next one or two digits provide additional information available in a broad subset of samples. Finally, trailing digits provide detail only rarely available.

The classification scheme for marital status illustrates the composite coding approach. Under the IPUMS-International design, shown in Table 3, the first digit of marital status has four categories: single, married/in union, separated/divorced/spouse absent, and widowed. This is the maximum number of categories consistently distinguishable across all samples in the database. The distinction between divorced and separated is not maintained in all samples, so these categories are combined in the fully comparable first digit of marital status. At the second digit, divorced and separated persons can be distinguished, as can formal marriages from consensual unions. The third and final digit differentiates among types of marriages (civil, religious, polygamous)— information only available for select countries.

The process of harmonizing variables starts with the documentation. The international dimension of the database requires careful attention to differing meanings of questions and responses, and involves comparing sometimes strikingly different systems of classification. The quality and quantity of variable documentation for discerning these meanings varies considerably across samples. At a minimum, all samples have two sources of variable-level information. Every sample has a dataset codebook that provides labels corresponding to the codes for each variable. But the labels are often insufficient to categorize a value in the context of all other countries in the database. For instance, the meaning of "unemployed" or even "married" varies by country. In addition to codebooks we always have access to the original census questionnaires themselves, which show the census wording on the forms and any pre-defined categories for responses. When in doubt we consider the census questionnaire the most basic source for discerning the meaning of variable categories, because it was most immediately in mind when the census questions were being answered. When the census forms are insufficient, the census enumerators' instructions usually provided the necessary information to discern the meaning of particular codes.

To make the harmonization process manageable, for most variables we separately perform a preliminary integration of each country's samples before turning to the final international integration. This focuses the compilation of information first on the time dimension,

Table 3. Coding Scheme and Category Availability for Marital Status

		Colombia				France					Kenya		Mexico				United States				Vietnam	
Code	Label	64	73	85	93	62	68	75	82	90	89	99	60	70	90	00	60	70	80	90	89	99
100	SINGLE/NEVER MARRIED	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	MARRIED/IN UNION																					
210	Married (not specified)	X	X	X	X	X	X	X	X	X	X	X	.	.	.	.	X	X	X	X	X	X
211	Civil	.	.	.	.	.	.	.	.	.	.	.	X	X	X	X	.	.	.	.	.	.
212	Religious	.	.	.	.	.	.	.	.	.	.	.	X	X	X	X	.	.	.	.	.	.
213	Civil and religious	.	.	.	.	.	.	.	.	.	.	.	X	X	X	X	.	.	.	.	.	.
214	Polygamous	.	.	.	.	.	.	.	.	.	X	X	.	.	.	.	.	.	.	.	.	.
220	Consensual union	X	X	X	X	.	.	.	.	.	.	.	X	X	X	X	.	.	.	.	.	.
	SEPARATED/DIVORCED/SPOUSE ABSENT																					
310	Separated or Divorced	.	X	X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
320	Separated	.	.	.	.	.	.	.	.	.	X	X	.	X	X	X	X	X	X	X	X	X
330	Divorced	.	.	.	.	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
340	Married, spouse absent (n.s.)	X	X	X	X	X	X	X	X	X	X	X	.	.	.	.	X	X	X	X	X	X
341	MSA, civil	.	.	.	.	.	.	.	.	.	.	.	X	X	X	X	.	.	.	.	.	.
342	MSA, religious	.	.	.	.	.	.	.	.	.	.	.	X	X	X	X	.	.	.	.	.	.
343	MSA, civil and religious	.	.	.	.	.	.	.	.	.	.	.	X	X	X	X	.	.	.	.	.	.
344	MSA, polygamous	.	.	.	.	.	.	.	.	.	X	X	.	.	.	.	.	.	.	.	.	.
350	Consensual union, spouse absent	X	X	X	X	.	.	.	.	.	.	.	X	X	X	X	.	.	.	.	.	.
400	WIDOWED	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
999	UNKNOWN/MISSING	.	X	X	X	.	.	.	.	.	X	X	X	X	X	X	.	.	.	.	X	X

Notes: Samples are identified by the last two-digits of their census year. An "X" indicates the category is available in that sample.

highlighting changes in census practice that might have been submerged in the wider international context.

In addition to pre-existing sample documentation such as questionnaires, the IPUMS-International project commissioned a series of topical essays from specialists in each country.

These were designed to provide insight on integration problems and possibilities from the perspective of persons with extensive experience with census data for a given country. Typically, these essays were written on clusters of variables, such as economic characteristics, education, or housing. The value of these contributions varied, but in some cases they provided the only information at our disposal on certain variables beyond what was printed on the census form itself. The essays also allowed the specialists to highlight what they considered to be major comparability or data quality issues, drawing from their own knowledge and experience.

The final step in harmonization requires considering all countries simultaneously to create an integrated variable coding scheme and corresponding documentation. We were largely on our own at this stage of the process. The work of our international partners did not encompass international variations. Nor were there alternative documents that could help greatly in this task of international integration. United Nations census recommendations were the closest thing to a standard, but they were meant as a guide to census-taking, not as a primer on integrating already-existing samples[3].

As always, with complex variables our intent is to make the first one or two digits comparable across all samples and add trailing detailed digits, as necessary, to retain variations available in some samples and not others. In contrast to our experience with IPUMS-USA, there were instances in which even at the most general level we did not make the categories fully comparable. For example, for the relationship variable, the 1999 census of Vietnam combined "other relatives and nonrelatives" in a single category, whereas every other sample made this critical distinction (ipums.org/international/codes/relate_codes.shtml). Full comparability in the first digit was achievable, but only at the cost of making the codes much more cumbersome for the

[3] For some of the most difficult variables—educational attainment, occupation and industry—there were international classifications that served as useful frameworks for integration (International Labor Office 1990; United Nations 1990; UNESCO 1997).

majority of samples, requiring users to go past the first digit of the variable to get this fundamental kinship information. In essence, we allowed some parallel, incompatible coding to persist in order not to inconvenience the majority of users with a clumsy coding structure. The price of increased convenience is the necessity for greater vigilance by the users. There is no rule for determining at what point such compromises are warranted, we simply had to exercise our judgment as census microdata users.

When no coding design can be devised that does not lose information or necessitate an unworkable classification scheme, we allow the loss of some detail in the integrated variables. To compensate, we keep a separate parallel unrecoded version of the variable to ensure no original detail is sacrificed.

Even when an elegant coding design is possible for a variable, in the international context the coding structure and labels may not be able to convey important comparability issues. Thus, the variable documentation has a substantial burden in IPUMS-International to provide warnings about drawing incorrect inferences from seemingly similar categories. We expect that feedback from users and our own experience with the data will suggest future edits to the variable descriptions and coding designs.

Harmonization of Employment Status – An Example

The IPUMS-International variable "employment status" illustrates the overall process of integration. We first determined the availability of the variable. Among the first-release countries, every sample except Mexico 1960 had a conceptually comparable variable, although it was called different things[4]. The actual work of integration began with an examination of the codebooks to discover the available variable categories in each sample. A partial listing of these categories is presented in Table 4.

Determining the meaning of the categories was the next step. Some appeared self-explanatory, but it was always necessary to examine the census questionnaires and the instructions given to the census enumerators. Table 5 shows the instructions that accompanied the employment status question in the 1999 Kenya census. Using the

[4] [4] Mexico 1960 has a related variable for number of days worked last week, but it lacks an explicit indication of unemployment.

Table 4. Code Availability for Employment Status, Selected Censuses

Colombia 1993		France 1975		Kenya 1999	
B	Missing	0	Inactifs autres que ceux cites ci dessous	B	N/A (under age 5)
1	Busca trabajo, tenia trabajo	1	Actifs ayant un emploi	0	[undocumented]
2	Busca trabajo, primera vez	3	Personnes sans emploi et en recherchant	1	Work pay profit
3	No trabajo pero tenia	4	Anciens actifs (retraites, retires des affaires…)	2	On sick leave
4	Trabajar	5	Etudiants et eleves nes avant le 1 janvier 1962	3	Family holding
5	Estudiar	6	Militaires du contingent	4	Family agricultural holding
6	Oficios hogar			5	Seeking work
7	Incapacitado			6	No work available
8	Jubilacion			7	Student
9	No response			8	Retired
0	Otra situacion			9	Incapacitated
				10	Home maker
				11	Other

Mexico 2000		United States		Vietnam 1989	
B	Missing	0	N/A	1	Worked 6 months and over
10	Trabajo	10	At work	2	Worked permanently less than 6 months
13	Trabaja se declara que busca trabajo	11	At work, public emergency	3	Worked temporarily less than 6 months
14	Trabaja se declara que es estudiante	12	Has job, not at work last week	4	Unemployed
15	Trabaja se dedica a los quehaceres del hogar	13	Armed forces	5	Student
16	Trabaja se declara que es jubil o pens	14	Armed forces	6	Household duties
18	Trabaja se declara que no trabaja	15	Armed forces, not at work last week	7	Invalid
19	Trabaja no se tiene informacion	20	UNEMPLOYED	8	Others
20	Tenia trabajo per no trabajo	21	Unemployed, experience worker		
30	Busca trabajo	22	Unemployed, new worker		
40	Estudiante	30	NILF		
50	Quehaceres del hogar	31	Housework		
60	Es jubilado o pensionado	32	Unable to work		
70	Esta incapacitado permanente para trabajar	33	School		
80	No trabaja	34	Other		
99	No especificado				

Table 5. Enumerator Instructions for Employment Status, Kenya 1999 Census

Column P30: Labor Force Participation
86. Column P30 contains a question on labor force participation during the LAST SEVEN DAYS PRECEDING THE CENSUS NIGHT and is asked of ALL PERSONS AGED 5 YEARS AND ABOVE

87. Ask all persons aged 5 years and above.
What was this person **MAINLY** doing during the last seven days proceeding the **CENSUS NIGHT?** What the respondent was **MAINLY** doing will denote the activity that occupied most of the respondent's time during the 7 days preceding the **CENSUS NIGHT**. The responses in column P30 are as follows:

Worked for Pay
Comprises persons who during the 7 days preceding the CENSUS NIGHT worked most of the time for wages, salaries, commissions, tips, contracts and paid in kind (especially in the rural areas where people who have rendered services may be paid using food or clothing).

On leave/sick leave
This group comprises all those with formal attachments to a job or business/enterprise but were not working during the reference period because they were sick or on holiday, seasonal workers, leave without pay, bad weather, etc. However, a person who is on leave such as a teacher but worked on family holding in the past 7 days preceding **CENSUS NIGHT** should be indicated as on leave.

Worked on Own/Family Business
This category comprises self-employed persons who worked on own business or persons who worked on family business for family gain. It includes "jua-kali" artisans, mechanics, traders in farm produce and family workers not on wage employment. Any member of the household working on the holding for pay will fall under code "01".

Worked on Own/Family Agricultural Holding
A holding in this case is the unit of land, farm or shamba which is owned or rented by the family and is used for purposes of cultivation or rearing livestock for subsistence. All the members of the household who are working on the holding without pay/profit will be coded "04" (i.e. working on Own/Family Agricultural Holding). Any member of the household working on the holding for pay will fall under code "01" (i.e. worked for pay).

Seeking Work
A person who in the 7 days preceding the **CENSUS NIGHT** was actively looking for work. This category should not include the under-employed (i.e those who have paid work but wish to leave for better opportunities). Persons who have no work at all and are looking for work will fall under this category. If a person is working on the family holding but is seeking work, he/she should be coded as "working on family holding" and not as "seeking work". This category should include only persons who are available full time for work and hence are' actively looking for it.

No Work Available
This is a person who is not working nor is looking for work because he/she is discouraged, but would usually take up a job when offered one.

Full-time Student
This is a person who spent most of his/her time in a regular educational institution (primary, secondary, college, university etc.) and hence not available for work. If, for instance, a student was on holiday during the 7 days preceding the **CENSUS NIGHT** and may have been engaged in gainful employment, he/she should be given the appropriate code "01".

Retired
This is a person who reports that during the 7 days preceding the **CENSUS NIGHT**, he/she was not engaged in any economic activity because he/she had retired either due to age, sickness or voluntarily. If a person has retired and is doing some work/business he/she should be coded appropriately, either as "01", "03" or "04". If he/she has retired and is seeking work he/she should be coded as "05".

Incapacitated
Is one who cannot work. Do not assume that all physically disabled persons cannot, work. For example, a blind person who is in wage employment will fall under category "01" and not "08". Similarly lame/crippled persons working on the family holding should fall under category "03" or "04". Please probe.

Homemaker
Is a person of either sex involved in household chores in his/her own home e.g. fetching water, cooking, babysitting, etc. who did not work for pay or profit nor sought work. These categories should not include house boys/girls who fall under category "01". If such a person worked on family holding they should be coded as "03" or "04" and not as "10". Please probe.

Other
This category Includes any other persons not mentioned above. You are to probe to find out whether unpaid family workers consider themselves as seeking work', etc. and code them accordingly. For example, if a young man helps his uncle to sell goods in a kiosk without receiving pay, probe whether he is 'seeking work' and code him thus; if he considers himself as working code him as "01".

For persons aged below 5 years leave column P30 blank. For respondents aged 5 years and above whose labor force participation status is not known or not stated, write "99".

Table 6. Translation Table for Employment Status

Harmonized Codes and Labels		Source Data Codes (selected samples)									
IPUMSI Code	IPUMSI Label	Col 1964	Col 1993	Fra 1962	Fra 1975	Ken 1999	Mex 1970	Mex 2000	US 1960	Viet 1989	Viet 1999
0000	N/A	*,5	B	*	B	BB	0	BB	00	B	B,1
	ACTIVE (In Labor Force)										
1000	EMPLOYED, not specified	1								1	
1100	At work		4	1	1	01	1	10	10		
1101	At work, and 'student'							14			
1102	At work, and 'housework'							15			
1103	At work, and 'seeking work'							13			
1104	At work, and 'retired'							16			
1105	At work, and 'no work'							18			
1106	At work, public emergency								11		
1107	At work, family holding, not specified										
1108	At work, family holding, not agricultural					03					
1109	At work, familiy holding, agricultural					04					
1110	Working and studying (France)										
1200	Have job, not at work last week		3			02		20	12		
1300	Armed forces								13		
1301	Armed forces, at work								14		
1302	Armed forces, not at work last week								15		
1303	Military trainee (France)			8	6						
2000	UNEMPLOYED, not specified	2			3	05	2	30	20		
2001	Unemployed (Vietnam)									4	5
2002	Worked less than 6 months, permanent job									2	
2003	Worked less than 6 months, temporary job									6	
2100	Unemployed, experience worker		1						21		
2101	Seeking work, worked less than 3 months			2							
2102	Seeking work, worked 3 to 6 months			3							
2103	Seeking work, worked 6 to 12 months			4							
2104	Seeking work, worked more than 1 year			5							
2105	Seeking work, experience unspecified			6							
2200	Unemployed, new worker		2	7					22		
3000	INACTIVE (Not in Labor Force)								30		
3100	Housework	3	6			10	3	50	31	6	2
3200	Unable to work/disabled	7	7			09		70	32	7	4
3300	In school	4	5	9	5	07		40	33	5	3
3400	Retirees and living on rent	8						60			
3401	Living on rent payments										
3402	Retirees/pensioners		8		4	08					
3500	Elderly	6									
3600	No work available/discouraged					06					
3700	Inactive, other reasons	9	0	0	0	11	4	80	34		6
9000	UNKNOWN/MISSING		9			00	9	99			9

Note: In the source data columns: a comma indicates more than one code was coded to the respective IPUMS-International value; an asterisk means programming logic was used; B indicates a blank in the source data.

documentation and coding information, we aligned the codes to equate comparable categories within each country individually.

The international integration began by determining the key distinctions maintained across countries, and consequently the maximum number of sustainable categories that could be imposed universally. For employment status, the critical information each census aimed to capture was participation in the labor force and unemployment.

The resulting IPUMS-International classification for employment status is shown in Table 6. The leftmost columns give the IPUMS codes and their labels, and the columns to the right show the original codes in each sample that correspond to the particular IPUMS code with which they are aligned on the same row.

The first digit of employment status has 3 categories that are largely comparable across all samples—employed, unemployed, and not in the labor force. (The "unknown and missing" codes will be allocated, and thus will disappear from later versions of the database.) After the first digit, national and temporal variations become evident. Among the unemployed, some samples distinguished between persons with past work experience (experienced unemployed) and persons seeking work for the first time (new workers). The number of categories distinguished among the inactive population varies widely among samples.

The final step was writing the variable documentation. This involved assessing what critical information was not fully conveyed by the coding structure for the variable or otherwise was of such importance that the user's attention should be drawn to it. One key aspect of employment status that required explanation was unemployment. The unemployed population is difficult to define consistently across countries. We attempted to apply U.N. and ILO standards in defining the unemployed as persons who are out of work and actively seeking a job. Some countries have relatively small paid-labor sectors and irregular labor markets, making unemployment comparisons difficult. Kenya identified persons who were not working simply because no work was available, explicitly referring to them as "discouraged" workers in the 1999 enumeration instructions (Table 5). In Kenyan census tabulations, these were considered unemployed, but in IPUMS-International they are coded as "not in the labor force" to maintain consistency with other countries.

A second major issue concerned the varying reference period for the employment status question. For most samples, employment status was reported with respect to the day of the

census or within a specified week prior to the census. For Vietnam, however, the reference period was the previous year - amounting to "usual employment status" over this period. These and other points are covered in the on-line variable description and comparability discussion for employment status. The final demonstration of variable harmonization is the ability to perform comparative analyses with the data. Figure 1 graphically presents a simple tabulation of the employment status variable. It gives female labor force participation rates by age for the 1990 round of censuses for each country in the first data release. The figure shows the markedly lower participation rates for the two Latin American countries, even in comparison to Kenya and Vietnam, which are actually quite similar to the two developed countries in the database. The figure is only a simple demonstration—it is for other researchers to say to what extent the differences are due to actual market labor allocation, treatment of household farms, retirement, marital status, cultural interpretations of work, or other possibilities. At the very least, the results are provocative and suggest genuine substantial international differences.

Figure 2 demonstrates the other dimension of the IPUMS-International data: time. The figure takes Mexico, the country with the lowest female labor force rates in Figure 1, and superimposes the rates for 1970 and 2000. The figure shows that in 1990 Mexico was in transition. By 2000 its rates were similar to those for 1993 Colombia, and the overall pattern began to show the telltale signs of a female population that was beginning not to exit the labor force upon getting married. A final way of organizing these results would track particular birth cohorts of Mexican women across the three censuses to separate cohort and period effects.

Constructed Variables

In almost all cases, census authorities collected data on households and relationships of individuals within households. With a few exceptions, family interrelationships are preserved in the microdata. IPUMS-International will create individual-level variables describing interrelationships among family members so that researchers can create specialized measures tailored to specific research topics, such as living arrangements of the aged or of single parents.

Figure 1. Female Labor Force Participation by Age by Country, ca. 1990

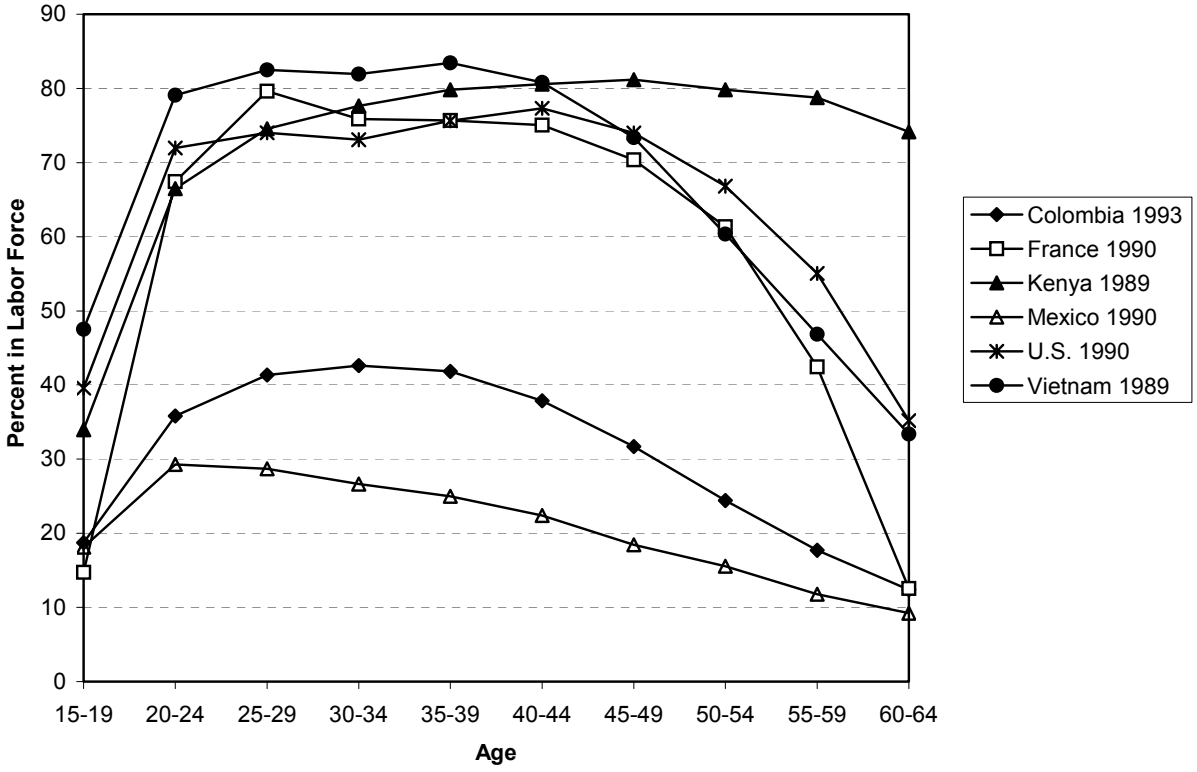
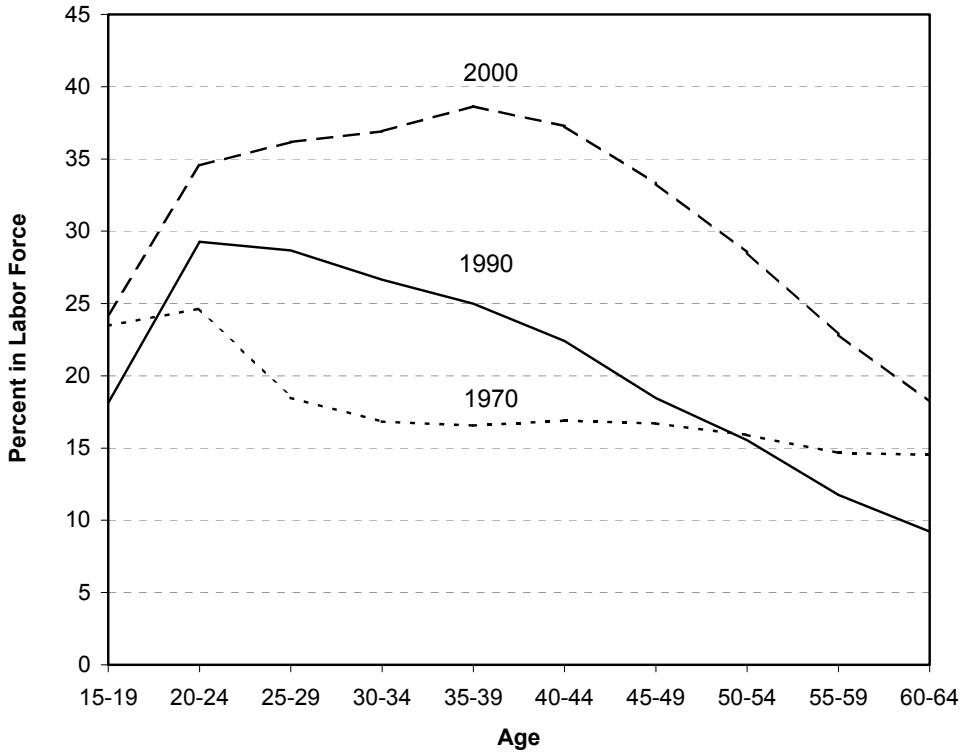


Figure 2. Female Labor Force Participation by Age, Mexico 1970-2000



Three pointer variables give the location within the household of each individual's mother, father, and spouse (or consensual partner). These pointer variables are among the greatest contributions we can make to the datasets. They allow users to easily attach characteristics of these kin to the records of other individuals and to create new family variables.

We are close to adding a first version of these pointers to the database. Because of both data and cultural differences, these family pointers are considerably harder to design and test than their counterparts that we built into the U.S. IPUMS. In all likelihood we will continue to adapt and improve them for the foreseeable future.

We will also construct several simple fully compatible variables describing family and household characteristics at the individual and household level. These indicators include those in IPUMS-USA: family membership, family size, number of own children, number of own children under five years old, and age of eldest and youngest own child.

Documentation

The creation of comprehensive integrated documentation is central to the project and is among its greatest challenges. For most users, the key documentation element is the detailed description of every variable, which includes universe definitions, frequency distributions, and variable codes[5]. The core variable description is supplemented by comparability discussions describing any deviations of particular censuses from the standard variable definition. The comparability discussions address differences over time and across countries. The variable pages also provide direct access to the wording of census questions, enumerator instructions, and facsimiles of census forms.

Where the documentation for the variables is concerned, there are competing goals and impulses. It is important to fully document each variable and all of its comparability issues, yet this risks deluging the user with so much information that the key points are not obvious. The solution is to provide only the most important information at the first level of documentation for a variable, but to provide further links from that page to deeper levels that provide more detailed information on comparability problems and variations. As we flesh out the variable discussions in the future, this multi-layered approach will be implemented.

[5] See http:/ipums.org/international/variables.shtml and click on the variable name.

We also provide English-language documentation on each of the samples included in the database. This integrated documentation will ultimately cover census enumeration procedures and instructions; definitions of households, dwellings, group quarters and other enumeration units; error correction and other post-enumeration processing; sample designs; census forms; and analyses of data quality, such as post-enumeration surveys[6].

As it grows, the data series will require the equivalent of thousands of pages of documentation. To manage this quantity of information, the web-based metadata access system will limit the scope of information to only those samples relevant to a given research project, as defined by the user. By constructing documentation pages dynamically, we can customize the documentation to the needs of particular researchers. For example, if a user selects censuses only for Colombia, s/he will only be offered information relevant to the Colombian samples. Comparability discussions will cover only the specific censuses selected by the user. Similarly, we will generate customized tables giving marginal frequency distributions restricted to the particular datasets chosen by the researcher. As we incorporate more samples into the database, this ability to filter out extraneous information will be critical, allowing us to provide documentation that devotes attention to subtle problems of comparability without overwhelming users with information they do not necessarily require.

Data Dissemination

Data access is an integral component of the project—in some sense the key component. Effective dissemination is essential if the data are to be widely used. Fortunately, the IPUMS-USA dissemination approach has proved highly successful, and it was directly adaptable to the needs of the international project.

The data extraction system presents the researcher with a sequence of screens that aim to keep the process as simple as possible. In the first screen the user chooses the samples—countries and census years—she or he wants. On the next page, the user is presented with a list of variables they can choose to include in the extract. The screen limits the choices to variables present in at least one of the samples chosen on the previous page. If the user

[6] The current sample descriptions are expressed in tabular form at http:/ipums.org/international/sample_designs.shtml. The questionnaires are accessible through the variable description pages and the "source materials" link on the main navigation bar.

checks any variable boxes for "case selection" they are presented with choices on the next screen. On the case selection page they can further limit their extract to only cases that have particular values for the chosen variables (e.g., females age 15 to 49). The final screen summarizes the choices, at which point the user can back up and revise them or instruct the system to create the extract. When the extract is complete, the user receives an email indicating the dataset is ready to be downloaded. All data files are in ASCII format, but the extract system creates SPSS, SAS, and Stata command files to facilitate reading the data into those statistical packages[7]. IPUMS-International is now developing second-generation data dissemination software. Among other improvements, we plan to add a feature allowing users to replicate data extracts used in published studies. The ability to replicate existing studies is essential to the scientific enterprise; it provides our fundamental means of understanding, evaluating, and building upon past research. In the new system we are developing, the codebook users receive with their extract will contain a recommended citation incorporating a unique number for the particular extract. We will encourage users to cite the extract number in their published work. Any authorized user will be permitted to specify that extract number and obtain a replica of the dataset. Thus, when scholars identify an extract number in their publications, readers of their work will be able to create and download an exact copy of the data used for the research.

Conclusion

The IPUMS harmonization strategy has proven flexible enough to accommodate the integration of data across broad spans of time (the United States for 1850-1990) and space (Colombia, France, Kenya, Mexico, the United States, and Vietnam). Our experience with IPUMS-International demonstrates that the composite coding strategy has the capacity to accommodate globally diverse marriage and familial structures—with relationships ranging from polygamous wives in Kenya to unmarried partners in France. As we peruse in advance other prospective datasets, we are already finding that most variable permutations around the world are accounted for in the existing coding designs. There will undoubtedly need to be further modifications and even complete overhauls of some variables as we

[7] The functionality of the extract system can be tested by clicking on "create an extract" on the main IPUMS-International page and logging in as "guest" using the password "guest".

encounter incompatible classifications. Nevertheless, we are confident that our overall approach will work, and that each additional sample will become easier to incorporate into the data series.

There can be no doubt that international integration is considerably more challenging than the integration of U.S. data alone. The number of samples and variables is greater, and the data quality and source documentation are more uneven. And the meaningful equivalence of some categories across countries and time is uncertain. Whether specific categories and concepts are actually not equivalent, however, is a question that should be addressed empirically. And it should be noted that most variables are fairly straightforward measures or relatively objective descriptors that are not especially vulnerable to differing cultural interpretations.

There are national and cultural differences that make for truly incongruent categories within variables. In IPUMS-USA, if a variable category was present in one year but not another, our method was to ask where that response would have been coded in the year that lacked that category. But that approach sometimes makes no sense in the international context. In different societies, that missing category may have no equivalent whatsoever.

The competing goals of full detail and usability produce starker choices for the international data series than for IPUMS-USA. It is possible with the international data to follow to absurdity the logics of complete integration and no loss of information. Coding to the least common denominator can lead to awkward variable classifications that would hinder most research. Instead—more often than with the U.S. data series—we will produce integrated variables that lose some detail while retaining the unrecoded sample- or country-specific "original" variables.

One of the unavoidable facts of the international data series versus IPUMS-USA is that it puts more burden on the user. Even when the dissemination system, coding schemes, and documentation are fully realized as we envision them, there will be ample opportunity for unwary researchers to make incorrect inferences and specious comparisons. IPUMS-International will always be more dependent than IPUMS-USA on documentation warnings, because the coding schemes cannot convey the shades of difference that arise. Not only do researchers need to read the documentation, they need to know something of the countries they are studying, at least as it pertains to their particular research topic.

The chance that researchers may misuse the data will always be there, but that is no excuse for not making access as open as possible. Concerns about making it easy to do bad research were voiced in response to the grant application that funded the original IPUMS project. Had that been the dominant opinion of the reviewers, a boon to the research community would never have been realized. It is the academic community's responsibility to ferret out bad research, as it always has been.

IPUMS-International is not simply a research tool, but a research project in its own right. The project was conceived by a research team consisting mainly of historians—but historians with a strong interest in the application of historical insights to policy issues today. The goal of the project is to encourage comparative research over space and time. The temporal emphasis has led us to privilege the inclusion of countries with multiple census samples, and that will remain a priority in the future. The difficulty for historical purposes is the paucity of microdata samples prior to computerization in the 1960s. There are, however, scattered older samples for various countries that might be added as the data series expands in the future.

Our guiding principle is to democratize access to data. With IPUMS-USA we expanded research opportunities for persons not affiliated with population centers to analyze the U.S. microdata. Anyone with an Internet connection and a desktop computer could work with the IPUMS. Now that democratization impulse is even more salient. Already, researchers from countries such as Kenya and Colombia are registering with IPUMS-International so that they can gain access to the data from their own countries. If things work out as we hope, IPUMS- International will expand to become the world archive for census microdata, giving researchers and policy-makers everywhere free and open access to these incomparable data sources.

References

International Labor Office. 1990. International Standard Classification of Occupations (ISCO-88). Geneva.

Kelly Hall, Patricia., Robert McCaa, and Gunnar Thorvaldsen, eds. 2000. Handbook of International Historical Microdata for Population Research. Minnesota Population Center: Minneapolis. (Updated microdata inventory available at www.ipums.org/international/iiinventory2.html.)

McCaa, Robert, and Steven Ruggles. 2002. The Census in Global Perspective and the Coming Microdata Revolution. In Vol. 13, Nordic Demography: Trends and Differentials, Scandinavian Population Studies, edited by J. Carling. Oslo: Unipub/Nordic Demographic Society, pp. 7-30.

Ruggles, Steven. 2000. "The Public Use Microdata Samples of the U.S. Census: Research Applications and Privacy Issues." A report of the Task Force on Census 2000, Minnesota Population Center and Inter-University Consortium for Political and Social Research Census 2000 Advisory Committee. (Available at: www.ipums.org/~census2000.)

Ruggles, Steven, and Matthew Sobek, et. al. 1997. Integrated Public Use Microdata Series: Version 2.0. Minneapolis: Historical Census Projects, University of Minnesota.

UNESCO. 1997. The International Standard Classification of Education (ISCED 1997). Paris.

United Nations. 1990. International Standard Industrial Classification of All Economic Activities (ISIC-88). United Nations Statistics Division. Department of Economic and Social Affairs, New York.

United Nations. 1998a. Principles and Recommendations for Population and Housing Censuses. United Nations Statistics Division. Department of Economic and Social Affairs, New York.

United Nations. 1998b. Recommendations for the 2000 Censuses of Population and Housing in the ECE Region. United Nations Economic Commission for Europe and Statistical Office of the European Communities. Statistical Standards and Studies, No. 49. New York and Geneva.