IPUMS-International

STEVEN RUGGLES
MIRIAM L. KING
DEBORAH LEVISON
ROBERT McCAA
MATTHEW SOBEK
Minnesota Population Center
University of Minnesota

Abstract. Integrated Public Use Microdata Series-International (IPUMS-International) is an effort to inventory, preserve, harmonize, and disseminate census microdata from around the world. IPUMS-International aims to convert census data from five continents into a uniformly coded and fully documented data series. Nearly all countries included in the data series contribute data from multiple census years. The project uses the same principles and methods that underlie the successful IPUMS-USA model. The data are distributed free of charge via a Web-based data extraction system. Use of the data is restricted to scholarly and educational purposes. The project is conceived as part of a larger enterprise to make the world's microdata available to researchers, encouraging cross-national research with a temporal dimension.

Keywords: census, demography, IPUMS, microdata, population

A vast quantity of raw individual-level census data for dozens of countries in the period since 1960 survives in machine-readable form. Most of these data, however, remain inaccessible to researchers. Integrated Public Use Microdata Series-International (IPUMS-International) is an effort to inventory, preserve, harmonize, and disseminate census microdata from around the world.

Unlike aggregated census tabulations, census microdata provide information about individual persons and households. This information enables researchers to design analyses tailored to their particular research questions. Other microdata sources—such as demographic and labor-force surveys—often offer greater subject coverage and detail than do census data, but no alternate source offers comparable sample density, chronological depth, and geographic coverage.

In the United States and Canada, census microdata have been available to researchers for almost 40 years and have become an indispensable component of social science infrastructure. For example, census microdata were the data source for 19 of the 51 U.S. and Canadian articles that appeared in the last two volumes of the journal *Demography* (2000 and 2001). Even though the United States has abundant high-quality survey data and the most recent census samples were over a decade old, U.S. census microdata were used three times as often as the next most popular data source. By contrast, during the same two years not a single article in *Demography* made use of census microdata from the developing world.

Virtually every country in the world created machine-readable census microdata in the course of conducting the censuses of the last 40 years. In many cases, however, these machine-readable files are endangered because of technological change and aging electronic media. Our first goal, therefore, was to identify and preserve surviving machine-readable data whenever possible. Even when microdata have been appropriately archived, however, they are usually inaccessible to researchers. A second goal, therefore, was to seek agreements from each national statistical agency to disseminate the data for scholarly research and education, subject to strict confidentiality protections.

IPUMS-International aims not only to make international census data available but also to make them readily usable. Even when census microdata are obtainable, comparison across countries or time periods is difficult because of inconsistencies between data sets and inadequate documentation of comparability issues. Consequently, comparative international research based on pooled census samples is rarely attempted. IPUMS-International reduces the barriers to international research by converting international census microdata into a uniform format, providing comprehensive documentation, and making the data available to researchers through a Web-based access system.

The project builds on our work to create a harmonized database of U.S. microdata between 1850 and 2000 (IPUMS-USA) (see Ruggles and Sobek 1997).[1] IPUMS-International extends the IPUMS model beyond the United States. Begun in 1999, the first phase of IPUMS-International includes microdata samples from 8 countries with broad geographical distribution. These data are being cleaned, harmonized, documented, and disseminated 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 (NIH). A project to add data for 16 additional Latin American countries has just been funded by NIH, and projects to add data from dozens of additional countries in Europe, Africa, and Asia are in the planning stages.

A preliminary database describing 48 million persons in six countries was released in May 2002. These data include the most commonly requested variables for 21 samples of the censuses of Colombia, France, Kenya, Mexico, the United States, and Vietnam between 1960 and 2000.

Table 1 shows the samples that are projected for inclusion 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. One goal of this diverse collection was to identify 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 both its inherent importance and the prospect for additional samples in the future.

In the final two years of the project, we will add all remaining variables contained in the various IPUMS-International data sets. These data sets will include 1982 China and five Brazilian samples from 1960 to 2001. Depending on time and resources, we may include Hungary, Ghana, and/or Spain as well.

IPUMS-International is a collaboration of the Minnesota Population Center (MPC) with dozens of partners, including national statistical offices, international archives, and research specialists in each country. A project of this complexity would have been impossible without the efforts of persons and institutions with knowledge of many specialized issues. The contributions of the UN Demographic Center for Latin America and the Caribbean (CELADE) and the UN Statistics Division were especially important. The national statistical offices also provided expertise, sample documentation, and in some cases prepared data samples for the project.

TABLE 1. IPUMS-International Samples

Country	Census year	% sample	Persons (000s)	Households (000s)
		2002 data release		
Colombia	1964	2	350	
	1973	10	1,989	350
	1985	10	2,643	571
	1993	10	3,274	788
France	1962	5	2,321	
	1968	5	2,488	
	1975	5	2,629	
	1982	5	2,714	
	1990	4.2	2,361	
Kenya	1989	5	1,074	225
	1999	5	1,410	318
Mexico	1960	1.5	503	
	1970	1	483	98
	1990	1	803	164
	2000	10.6	10,099	2,312
United States	1960	1	1,800	579
	1970	1[a]	2,030	744
	1980	1[a]	2,267	942
	1990	1[a]	2,500	1,106
Vietnam	1989	5	2,627	534
	1999	3	2,368	534
		2004 data release		
Brazil	1960	1[a]	914	
	1970	1[a]	1,105	
	1980	3[a]	3,526	
	1991	2.7[a]		
	2001			
China	1982	0.1	1,002	242
United States	2000	1 and 5	16,884	6,954

Note. Possible additional samples include Ghana 1984, 2000; Hungary 1980, 1990, 2001; Spain 1981, 1991, 2001. The French data sets currently do not include households, but that will be rectified in 2003.

[a]By March 2004, sample densities will be increased to 6% for U.S. censuses of 1970, 1980, and 1990. Sample densities for Brazil may increase.

Preservation

A central goal of IPUMS-International is to create an inventory of surviving census microdata and documentation. We are collecting enumerator instructions, census forms, codebooks, studies of data quality, and any other ancillary documentation we can locate for all countries that will allow us access to this information. Whenever feasible, we have obtained copies of microdata as well as documentation.

The microdata inventory meets several needs. First, it constitutes an important resource in its own right for researchers and data archivists. Second, the inventory underpins the design of the IPUMS-International database, allowing us to design the system to accommodate future expansion by taking into account the range of variation in census content and concepts around the world. Third, the microdata inventory helps us identify additional census and survey resources that should be preserved.

The first edition of our inventory of machine-readable census microdata was published in the *Handbook of International Historical Microdata for Population Research*, edited by Patricia Kelly Hall, Robert McCaa, and Gunnar

Thorvaldsen (2000), which received the 2001 best book prize from the American Association for History and Computing. An up-to-date listing of the inventory is available at the project Web site (http://ipums.org/international).

Of some 250 historical census microdata sets we have identified, almost 100 are being preserved under the direct auspices of the IPUMS-International initiative. Our preservation efforts extend to all machine-readable data sets endangered because of technological obsolescence, even when there is little likelihood that the data will be made available to researchers. Much of this preservation effort has been carried out via CELADE, which has large historical data holdings for Latin America. We also worked with the East-West Center in Hawaii, which has extensive collections of Asian and Pacific censuses.

Complete and comprehensive metadata are essential for the use and interpretation of census microdata. The MPC took an important step toward the preservation of census metadata when it acquired the Historical Archive of census documentation (approximately 150 linear feet of printed materials) from the UN Statistics Division in 2000. We have scanned the census enumeration forms in the collection to create electronic images of over 400 census forms from the 1950s to the present. Products from this acquisition are available as *World Census Questionnaires: First Edition*, available in PDF format through the IPUMS-International Web site.

Confidentiality

Privacy considerations are a worldwide concern, but many official census agencies are beginning to recognize that legal, organizational, and technical procedures can safely protect public use microdata files from misuse. In contrast to the United States, where census microdata samples are available to the public without any restrictions, the international samples require users to satisfy the confidentiality requirements of national statistical agencies. To meet these needs, we use two strategies for safeguarding the confidentiality of the microdata: confidentiality agreements and statistical disclosure protections. Used in combination, these approaches minimize the potential risk of disclosure without seriously compromising scientific use of the data.

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 do the following:

- Maintain the confidentiality of persons, households, and other entities. Any attempt to ascertain the identity of persons or households from the microdata is prohibited. Alleging that a person or household has been identified is also prohibited.
- Implement security measures to prevent unauthorized access to census microdata. Under IPUMS-International agreements with collaborating agencies, redistribution of the data to third parties is prohibited.
- Use the microdata for the exclusive purposes of 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 IPUMS-International, which will in turn pass on the information 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. If 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. MPC employees who work with the census microdata also sign agreements to respect the confidentiality of the data.

Technical safeguards supplement these institutional controls. We are working with each country's statistical office to minimize the risk of disclosing respondent information. The details of the confidentiality protections vary across countries, but names and detailed geographic information are suppressed in all cases. In addition, we use a variety of other procedures to enhance confidentiality protection. These include 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. These procedures are designed to extend this record.

Data-Set Reformatting, Cleaning, and Data Allocation

We carry out a systematic program of data reformatting and cleaning for each data set. The original microdata samples are preserved in a wide variety of formats ranging from separate household and person files to multilevel hierarchical files. This welter of file structures must be standardized

before we can carry out harmonization routines. Separate programs transform each data set into a hierarchical format containing household and person records. Any additional levels of information above the household, such as dwellings, are recorded on the household record. This data restructuring not only provides regularized input for the harmonization routines but also uncovers errors that could not be identified from a simple examination of data frequencies. Restructuring is thus an integral aspect of data cleaning. The data reformatting procedures are described in detail in Esteve and Sobek's article on pp. 66–79 in Part Two of this issue.

We have developed a battery of tests to ensure data soundness. Whereas most of the data sets are generally of high quality, they often have internal inconsistencies. Among the conditions 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. When 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 that are then replaced with allocated values 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 on the basis of the sex of a spouse.

When missing or inconsistent items cannot be resolved through logical computer editing, we turn to hot-deck probabilistic allocation procedures modeled on those of the U.S. Census Bureau (Ruggles and Sobek 1997, vol. 3). Such allocation procedures increase the reliability of sample estimates and simplify usability. 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. If a perfectly matched donor record cannot be found, the record that meets the largest number of criteria is used. The donated value is then subjected to consistency checks and is rejected if unsuitable. A data-quality flag identifies any allocated or edited data items.[2]

Variable Harmonization

The international census samples employ differing numeric classification systems, and reconciliation of these codes is a major part of the work of IPUMS-International. The goal is to create variables in which codes mean the same thing across all data sets.

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 that 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 those cases, we construct composite coding schemes. The first one or two digits of the code provide 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.

Most data transformations are simple recodes of one value into another. To carry out the recoding, we develop data transformation matrices for each variable that provide information on the location of the original variable in each sample, each original data value, and each new standardized data value. In many instances, it is necessary to use information from more than one variable in the original census to construct a new compatible variable. For example, one might need information on both province and subdistrict to identify a metropolitan area. Data transformation matrices can often handle such complex transformations, but in other cases we resort to customized programming solutions.

For some variables, it is impossible to design a fully compatible classification, even with the use of composite coding schemes. In such instances, we create whatever least-common-denominator classifications are applicable across multiple data sets. To ensure that no information is sacrificed, we include unrecoded versions of the original variables.

Constructed Variables

In addition to recoding variables to maximize comparability, we carry out further processing to enhance usability. Some procedures—such as the addition of compatible variables on serial number, census year, country code, size of household, and case weights—are straightforward. Others are more complicated; some examples follow.

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 creates

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. 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 make to the data sets. They allow users to easily attach characteristics of these kin to the records of other individuals and to create new family variables. Because of both data and cultural differences, these family pointers are considerably harder to design and test than their counterparts that we built into IPUMS-USA. We expect to adapt and improve them over the life of the project.

We also construct several simple fully compatible variables describing family and household characteristics at the individual and household level. These indicators include family membership, family size, number of own children, number of own children under 5 years of age, 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.[3] 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.

The documentation for the variables must balance competing goals. It is important to fully document each variable and all of its comparability issues, yet such documentation risks deluging the user with so much information that key inconsistencies are not obvious. The solution is to present only the most important information at the first level of documentation for a variable and to provide further links from that page to deeper levels that provide more detailed information on comparability problems and variations.

We also provide English-language ancillary documentation on each of the samples included in the database. This documentation covers census enumeration procedures and instructions; definitions of households, dwellings, group quarters, and other enumeration units; error correction and other postenumeration processing; sample designs; census forms; and analyses of data quality, such as postenumeration surveys.[4] The original-language versions of the underlying documents will also be available. The documentation will describe all data transformations that we perform on the original data to generate the integrated database.

Ultimately, 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 be offered information relevant only 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 data sets chosen by the researcher. As we incorporate more samples into the database, this ability to filter out extraneous information will be increasingly important, allowing us to provide documentation that devotes attention to subtle problems of comparability without overwhelming users with information they do not require.

Data Dissemination

Effective dissemination is essential if the data are to be widely used. We have adapted the successful IPUMS-USA dissemination approach to the needs of the international project. A Web-based data access system allows users to custom-design data extracts containing only the samples, variables, and cases required for their research. The data access system is fully integrated with the variable and sample documentation so researchers can make informed decisions as they define their data sets.

Once users have been registered, the data access system presents them with a sequence of screens to guide them in the design of their data set. In the first screen, users choose the samples—countries and census years—they want. On the next page, users are 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 users check 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 data set to those cases that have particular values for the chosen variables (e.g., females between the ages of 15 and 49). The final screen summarizes the choices and allows users to make revisions if necessary. When the extract is complete, the user receives an e-mail indicating that the data set 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.[5]

The extraction engine is designed to take advantage of the hierarchical structure of census data. We offer researchers the option of rectangular or hierarchical output files and allow users to select households or families on the basis of individual-level characteristics. Future versions of

the IPUMS-International data access system will add the following two additional features to make it easier for researchers to exploit the hierarchical structure of the data:

1. A procedure for attaching characteristics of household heads, family heads, spouses, own mothers, and own fathers to each individual's record. For example, the system will allow analysts of marriage to create new variables describing spouse's age or birthplace.

2. A procedure for counting the number of persons within each household, family, or own children of each parent that have a combination of up to four characteristics. For example, the data access system will be able to count the number of teenage daughters in the labor force for each mother with coresident children. The system will also sum numeric characteristics (e.g., income) across households, families, or own children.

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. As noted, we have already received funding to expand the project to cover virtually all of Latin America. We have also been scouring the rest of the world for opportunities to obtain and disseminate microdata samples. Those openings are proving to be much greater than we supposed when we first proposed the IPUMS-International project. 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 incorporate the data into IPUMS-International. Most surprising has been the number of positive reactions from countries with reputations for restrictive or nonexistent access to microdata.

As the database grows, 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.

Conclusion

The IPUMS harmonization strategy has proven flexible enough to accommodate variations across broad spans of time and space. The coding strategy has the capacity to accommodate global diversity in such areas as marriage and familial structures—with relationships ranging from polygamous wives in Kenya to unmarried partners in France. As we examine more and more data sets, we are finding that most permutations around the world are accounted for in the existing coding designs.

International harmonization is considerably more challenging than the harmonization of U.S. data alone. The number of samples and variables is greater, and the data quality and source documentation are more uneven. Although we will undoubtedly encounter new problems as we proceed, we are confident that our overall approach will work and that each additional sample will become easier to incorporate into the data series.

Our central goal is to democratize access to data. IPUMS-International data are available without cost to all scholars with a relevant research or educational project, which is a significant contribution to the data resources of the social science community. We can look forward to a harmonized database with hundreds of millions of records representing the majority of the world's population. It will not only open new opportunities for comparative cross-national research, but it will also provide researchers in many countries with research tools more powerful than they have ever had before.

NOTES

IPUMS-International is funded by a social science infrastructure grant from the National Science Foundation (SBR 9907416), Steven Ruggles, principal investigator. The procurement and development of Colombian samples is funded by the National Institutes of Health (R01 HD037508), Robert McCaa, principal investigator.

1. See http://ipums.org/usa.
2. Most U.S. Census Bureau allocation procedures were never published, but they are available in the IPUMS documentation, as are the procedures we developed for the U.S. censuses of 1850–1920 as part of the IPUMS-USA project; see Ruggles and Sobek 1997, vol. 3 (http://ipums.org/usa).
3. See http://ipums.org/international/variables.shtml; then click on the variable name.
4. 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.
5. The functionality of the extract system can be tested if one clicks on "create an extract" on the main IPUMS-International page and logs in as "guest" using the password *guest*.

REFERENCES

Hall, P. K., R. McCaa, and G. Thorvaldsen. 2000. *Handbook of international historical microdata for population research.* Minneapolis: Minnesota Population Center.

McCaa, R., and S. Ruggles. 2002. The census in global perspective and the coming microdata revolution. In *Nordic demography: Trends and differentials, Scandinavian population studies*, vol. 13, edited by J. Carling, 7–30. Oslo: Unipub/Nordic Demographic Society.

Ruggles, S. 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/~census 2000).

Ruggles, S., and M. Sobek et al. 1997. *Integrated Public Use Microdata Series: Version 2.0.* Minneapolis: Historical Census Project, University of Minnesota.