20 IPUMS-International

A Global Project to Preserve Machine-Readable Census Microdata and Make Them Usable

Robert McCaa
Steven Ruggles

Introduction

The IPUMS-International project has four goals:

1. Inventory machine readable census microdata.
2. Preserve census microdatasets identified as at-risk.

And for those countries in the IPUMS-International consortium:

3. Create an integrated international census database with a harmonized system of concepts, variables and codes, incorporating both historical and contemporary microdata of individuals, households and dwellings.
4. Disseminate integrated microdata via the internet, using techniques similar to the IPUMS-USA web-based system http://www.ipums.org.

Robert McCaa is Professor of History at the University of Minnesota and a principal investigator on the IPUMS-International project at the Minnesota Population Center.

The IPUMS-International project is supported by a grant from the National Science Foundation. To learn more about this data collection and preservation project, or to provide information to update the Inventory of Known Microdata, please visit the IPUMS-International web page at: http://www.ipums.org.

The IPUMS-International project began in October 1999, with a five year grant from the National Science Foundation of the United States. A three year grant from the National Institutes of Health was awarded in May 2000 for the integration of Colombian census microdata.

Historical census microdata for Austria, Canada, Norway, Great Britain, and Argentina will be included in the project as well as those for the United States. Elsewhere in this volume specific chapters are devoted to several of these historical census microdata projects and will not be discussed further here (but see Chapters 2, 3, 8, 12 and 16).

Contemporary microdata for China, France, Great Britain, Hungary, Spain, Colombia, Mexico, and Brazil will be integrated into the database as well as those for the United States (see Chapters 4, 5, 7, 8 and 10). Negotiations are underway with statistical offices in a number of other countries (see below the section on "data expansion").

Source Material

The first task of the IPUMS-International project is to inventory census microdata currently known to exist and to preserve those datasets identified as at risk. The Inventory of Known Microdata in Section III presents the results of first efforts to identify existing microdata. Of 193 extant data sets tallied in the Inventory, nearly one hundred are being preserved under the auspices of this project. We suspect that another hundred are still in existence, but these remain unverified until physical existence is confirmed by means of a count of the actual number of currently surviving records.

The second task of the project is to make census microdata for selected countries usable (Table 20-1). Although large machine-readable census microdata exist for many countries, public access to these data is restricted in virtually every case. Countries that join the IPUMS-International integration consortium agree—given all appropriate privacy and confidentiality safe-guards—to permit public access to samples of their census microdata.

Complete and comprehensive metadata are essential to the success of the project. For every country where census

Table 20-1. Countries in the International Integration Database (as of June 19, 2000)

	Census Years
Argentina	1869, 1895
Austria	1961, 1971 1981 1991, 2001
Brazil	1960, 1970, 1980, 1991, 2001
Canada	1871, 1901
Colombia	1964, 1973, 1985, 1993, 2000
China	1982, 1990, 2000
France	1962, 1968, 1975, 1982, 1990
Great Britain	1851, 1881, 1961*, 1971*, 1981*, 1991, 2001
Hungary*	1970, 1980, 1990, 2001
Mexico	1960, 1970, 1990, 2000
Norway	1801, 1865, 1875, 1891, 1900, 1910, 1920
Spain	1970*, 1981, 1991, 2001
United States	Decennial 1850-2000 (except 1890)

*currently under review

microdata currently exist, whether they may be integrated into the international database or not, we seek to preserve four types of documentation for each dataset: codebooks, original enumeration schedules, enumerator instruction booklets, and data processing instructions. For the most recent census microdata these materials may be published, indeed in a single volume. An excellent model is the Hungarian Central Statistical Office's *1990 Population and Housing Census: Summary Report on the Data Collection and Processing*. (Budapest, 1995). For earlier censuses, some of these materials may be elusive, existing only in archival form, often as typescripts with, at times, only a single surviving copy.

Procedural History

The goal of the IPUMS-International project is not simply to make international microdata available; it will also make them

usable. Even in the few cases where microdata are already available, comparison across countries or time periods is challenging owing to inconsistencies between datasets and inadequate documentation of comparability problems (Domeschke and Goyer, *Handbook of National Population Censuses*). Because of this, comparative international research based on pooled microdata is rarely attempted. The IPUMS[i] project will reduce the barriers to international research by preserving datasets and making them freely available, converting them into a uniform format, providing comprehensive documentation, and developing new web-based tools for disseminating the microdata and documentation.

The integration project entails two complementary tasks: first, the collection of data that will support broad-based investigations in the social and behavioral sciences; and second, the creation of a system incorporating innovative capabilities for worldwide web-based access to both metadata and microdata.

The integration project is composed of four interrelated elements. The first is planning and design. The international dimension of the database poses new design challenges, since it must accommodate variations in census design and cultural concepts. The starting point for developing an integrated design must be the standard classification schemes in the field of international population censuses, including, but not limited to, the following:

United Nations Statistical Division. *Principles and Recommendations for Population and Housing Censuses* (1998).

UNESCO. *The International Standard Classification of Education (ESCED 1997).*

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

United Nations Statistical Division. *International Standard Industrial Classification of All Economic Activities.*

The basic design goals remain the same as in the IPUMS-USA: the international system should simplify use of the data

while losing no meaningful information except where necessary to preserve respondent confidentiality.

The second element, microdata conversion, falls into two categories. For some countries, such as China, France, Mexico and Brazil, the project will incorporate already-existing public-use samples. For other countries, no public-use census files presently exist (e.g., Spain and Austria, and for microdata prior to the 1990s, Colombia, Great Britain, and Hungary). In these instances, new samples will be drawn from surviving census tapes using techniques to ensure that respondent confidentiality is preserved. These data files are often not publicly documented and require extensive assistance from the statistical offices and experts of each country to assure their correct interpretation.

The third element, the development of metadata, is central to the project and poses even greater challenges than the microdata. The documentation is not confined to codebooks, census questionnaires and enumerator instructions. As with the IPUMS-USA, a wide variety of ancillary information will be provided to aid in the interpretation of the data, including full detail on sample designs and sampling errors, procedural histories of each dataset, full documentation of error correction and other post-enumeration processing, and analyses of data quality.

The final element of the project is the creation of an integrated data access system to distribute both the data and the documentation on the Internet. With the IPUMS-International access system users will extract customized subsets of both data and documentation tailored to their particular research questions (unlike the IPUMS-USA system, where the entire documentation system is provided to the user, regardless of the data requested). The IPUMS-International system will consist of a set of tools for navigating the mass of documentation, defining datasets, and constructing customized variables. Given the large number of variables and samples, the documentation will be so unwieldy as to be virtually unusable in printed form. Accordingly, the project will develop software that will construct electronic documentation customized for the needs of each user.

Electronic Formats

In addition to ASCII text files, the project plans to disseminate microdata as pre-constructed SPSS and SAS system files. Documentation will be available in hypertext form.

Variable Availability

Variable design often influences the analytical strategies adopted by researchers, and we must therefore develop our plans with care. We have two competing goals. On one hand, we want to keep the variables simple and easy to use for comparisons across time and space. This requires that we provide the lowest common denominator of detail that is fully comparable, with underlying complexities transparent to the user. On the other hand, we must retain all meaningful detail in each sample, even when it is unique to a single dataset.

We will employ several strategies to achieve these competing goals. In some cases, the original variables are compatible and their recoding into a common classification is straightforward. The documentation will note any subtle distinctions a user should be aware of when making comparisons. 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 will construct composite coding schemes. The first one or two digits of the code will 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. The data access system will guide researchers to use only the level of detail appropriate for the particular cross-national or cross-temporal comparisons they are making. All data from the original enumerations will nevertheless be available to researchers who wish to use it.

In some cases, incompatibilities across samples are so great that the composite coding scheme is significantly more cumbersome than the original variable coding design. In these

cases, we will develop alternate versions of the variables suitable for particular comparisons across time and space. The data access system will recommend the most appropriate version of each variable to researchers based on user profile and the particular combination of datasets they are using. We anticipate that this approach will be needed more often in the international context than it was in the construction of the IPUMS-USA. Where feasible, we will base our coding designs on United Nations coding systems. For geographic variables, we will generally conform to the standard of the country.

Most data transformations are simple recodes of one value into another. As in the case of the IPUMS-USA, we will develop data transformation matrices for each variable which provide information on the location of the original variable in each sample, each original data value, and each new standardized data value. These matrices will be maintained in a standard relational database. The actual recoding operations, however, will be carried out with a C program operating as a sequential batch process, since that is the most efficient approach with respect to both storage and speed. In many instances, it is necessary to use information from more than one variable in the original 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 sometimes handle such complex transformations, but in other cases we will have to resort to customized programming solutions.

One of the greatest contributions of the IPUMS-USA to the original U.S. census files was the creation of family interrelationship variables in all years. We will construct similar variables for the international database. A system of logical rules identifies the record number within each household of every individual's mother, father, or spouse, if they were present in the household. These "pointer" variables allow users to attach the characteristics of these kin or to construct measures of fertility and family composition. For example, use of the spouse pointer variable makes it easy for users to identify spouse's income for each married person in the census. Because of variations across countries in the information available for identifying family interrelationships and in the cultural meaning

of marriage (e.g., the high frequency of consensual unions in Latin America and of cohabitation in Scandinavia), we plan to revise the logic of the family interrelationship variables for the international database.

We will also construct a wide variety of fully compatible variables describing family and household characteristics at the individual and household level. Some of these tools—such as family and subfamily membership, family and subfamily size, and number of own children—are incorporated in the existing IPUMS. For the new database, we will design new constructed variables to describe household and family composition in ways that reflect the diversity of family forms across countries.

Confidentiality Provisions

All accessible census microdata files are designed to protect the confidentiality of individuals. Countries have different standards, but in all cases names and detailed geographic information are suppressed and top-codes are imposed on variables such as income that might identify specific persons. Some countries take additional steps, such as "blurring" a small percentage of geographic information or randomizing the sequence of cases so that detailed geography cannot be inferred from file position.

Many datasets we will be working with will already have been subjected to confidentiality procedures by the national agency that created the files, and in these cases we will not need to take any additional steps. In other cases, however, we will be working with the original 100 percent machine-readable census returns, from which we will draw a nationally representative sample of specified density. In such cases, we will work closely with each country's statistical office to ensure full confidentiality of all files before they are made public. We will work to develop new methods to maximize the available detail while maintaining full confidentiality.

The IPUMS-International project distributes integrated microdata of individuals and households only by agreement of the corresponding national statistical offices and under the strictest of confidence. Before data may be distributed to an individual researcher, an electronic license agreement must be

signed and approved. To gain access to the data, researchers must agree to the following:

1. Recognize the copyright of the corresponding national statistical agency.
2. Use the microdata for the exclusive purposes of teaching, academic research and publishing, and not for any other purposes without the explicit written approval, in advance, of the corresponding national statistical authorities. Researchers must explicit agree to not use microdata acquired for the pursuit of any commercial or income-generating venture either privately, or otherwise. Please note that the corresponding national statistical authorities may at their discretion approve use for commercial purposes, but not the IPUMS-International project.
3. Maintain the absolute confidentiality of persons and households. Please note that any attempt to ascertain the identity of persons or households from the microdata is strictly prohibited. Alleging that a person or household has been identified in these data is also prohibited.
4. Implement security measures to prevent unauthorized access to microdata acquired.

Penalties for violating the agreement include revocation of the license, recall of all microdata acquired, filing of a motion of censure to the appropriate professional organizations, and civil prosecution under the relevant national or international statutes.

Research Possibilities

The potential list of topics that can be addressed with these data is far too long to discuss within the space constraints of this chapter. Among key research areas are economic development, poverty and inequality, industrial and occupational structure, household and family composition, the household economy, female labor force participation, employment patterns, population growth, urbanization, internal migration, immigration, nuptiality, fertility, and education. In each case, these topics can be studied across countries and across

census years. Analysts of immigration to the United States will be able to compare the characteristics of newcomers with those they left behind. Students of African, British, Italian, or Hispanic diasporas will be able to compare the characteristics of people with the same ethnic origin in a wide variety of countries, and to assess how those characteristics have changed over time in each country. Researchers working on the impact of NAFTA will be able to carry out multivariate analyses spanning the last three decades of the twentieth century using pooled microdata from Canada, Mexico, and the United States.

Data Expansion

While the first phase of the project includes only twelve countries, as additional microdata become available, we will seek funds to incorporate these into the international access system. We have six criteria for inclusion of additional international microdata into the database:

1. Public accessibility. In no case will the IPUMS-International project incorporate data unless it can be made available at no cost to certified academic researchers who are willing to sign a nondisclosure agreement. In most cases, the conditions of use will be only slightly more restrictive than they are for existing IPUMS data. For each country, we will draw up an agreement to allow the microdata to be distributed without charge.

2. Data quality. All samples included in the database must meet certain minimum criteria of data quality. They must have a net undercount of no more than 12%, as determined either through a post-enumeration survey or by demographic estimation techniques. In addition, IPUMS[i] expert consultants will review documentation relating to enumeration methods and the procedural history of each census to detect significant flaws in methods of data collection and processing.

3. Size. The costs associated with incorporating each additional sample do not vary greatly with size. To maximize cost-effectiveness, we will not normally

incorporate a sample unless there are at least 100,000 cases available for analysis.

4. Availability of key variables. All samples must provide information on individuals grouped into households, families, and/or dwellings.[1] At a minimum, each census must provide information on age, sex, marital status, occupation, and birthplace. In most cases, we will also require family relationships, education, and employment status.

5. Chronological depth. In all cases we will require a minimum of two available census years, and we prefer countries with four or more.

6. Cooperation of experts and availability of documentation. We will only include countries in which we can secure the assistance of national statistical agencies or other data experts, and where we can obtain adequate documentation specifying the details of creation of the microdata.

These criteria will not be the only factors we consider. In some cases we may bend the criteria on sample size, chronological depth, data quality, or key variables for samples with exceptional intrinsic interest. For example, we want to ensure geographic diversity of the database, and this may involve compromises. Moreover, we will give high priority to samples that are currently unavailable to scholars in any form.

[1] In the case of the United Kingdom and Canada we will include the non-hierarchical individual samples as well as the "family" samples even though they do not meet this criterion, because doing so will involve virtually no additional expense and will provide useful additional cases and geographic detail essential for some applications. For the 1960-round of census microdata for Mexico, Brazil and Colombia—indeed, for Latin American samples from the 1960s—only non-hierarchical samples of individuals were created. These will be considered on a case-by-case basis.

Bibliography

Domschke, Eliane and Doreen S. Goyer. (1986). *The Handbook of National Population Censuses: Africa and Asia*. Westport, CN.

Hungarian Central Statistical Office. (1995). *1990 Population and Housing Census: Summary Report on the Data Collection and Processing*. Budapest.

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

United Nations Department of Economic and Social Affairs Statistical Division. *Principles and Recommendations for Population and Housing Censuses.* New York.

_____. (1990). *International Standard Industrial Classification of All Economic Activities.* New York.

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

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