

**IPUMS-INTERNATIONAL HARMONIZED CENSUS MICRODATA EXTRACT SYSTEM:
USERS AND USES, MAY 2002-JANUARY 2005**

Robert McCaa, Steven Ruggles, and Matt Sobek
University of Minnesota Population Center

Research for this paper was funded in part by the
National Science Foundation of the United States, grant SBR-9908380.

Abstract. Census microdata are an invaluable resource for social science and policy research. Until recently National Statistical Institutes permitted little access to these data. This paper reports on users and uses of the IPUMS-International database (www.ipums.org/international). The project is a global collaboratory to anonymize, harmonize and provide access on a restricted basis to integrated microdata extracts of census samples. Custom-tailored extracts and documentation are delivered, at no charge via the Internet, but access is restricted to bona fide scientists with demonstrated research needs who agree to abide by the conditions of use license. This paper summarizes information on the users and uses of the database during the first 33 months of operation.

Introduction. Census *microdata* provide information about individual persons and often families, households, and dwellings, usually in the form of one or more records per case, each consisting of a series of variables. Typical census microdata variables for person records include age, sex, marital status, family relationship, place of birth, educational attainment, employment status, etc. are. Microdata are exceedingly useful because they allow researchers to interrelate any desired set of population and housing characteristics (Dale, Fieldhouse and Holsworth, 2000). Remarkably, the United Nations Statistics Division has long remained silent with respect to the use of census microdata. For example, the Principles and Recommendations for the 2000 round of population and housing censuses (UNSD 1998) offers little advice preservation of or access to microdata. Nevertheless, the flexibility offered by microdata is essential for comparative research because aggregate tabulations produced by national statistical offices are usually not comparable across time or between countries. In the few countries where census microdata covering multiple census years have been easily available to researchers, these data are the most widely-used source for the study of large-scale economic and demographic transformations (McCaa and Ruggles, 2002).

See Appendix Table 1

The IPUMS-International project is a global consortium to harmonize and disseminate high-density census microdata samples (Ruggles et. al. 2003). Begun in 1999 with funding provided by the National Institutes of Health and the National Science Foundation of the United States, to date the initiative enjoys the endorsement of official statistical agencies in more than fifty countries, encompassing more than half the world's population (Appendix Table 1). In May 2002, the first phase of integrated census microdata for Colombia (1964-1993), France (1962-1990), Kenya (1989-1999), Mexico (1960-2000), the United States (1960-2000), and Vietnam (1989-1999) were made available to researchers, followed by China (1982) in 2003 and Brazil (1960, 1970, 1980, 1991, 2000) in 2004.

Including the data for Brazil, the IPUMS-International website offers some 120 million person records consisting of more than 100 variables from 28 samples (Appendix Table 2). Over the next five years, thanks to sustained funding by the National Science Foundation and the National Institutes of Health, microdata for some three dozen countries will be added through regional initiatives in Europe, Latin America, Africa, Asia and the Pacific. We expect that Asian researchers will constitute the second largest group of users, after the United States, once additional samples from Asian countries are in the database. Following endorsement of the project's memorandum of understanding, complete microdata and documentation have already been received from the statistical offices of the Philippines, Malaysia, Mongolia, Cambodia, the Palestinian Authority, and Israel. Negotiations are underway with the census authorities of India, Indonesia, Pakistan, Bangladesh, and a number of others in the region.

See Appendix Table 2

Dissemination (“Extracts”). Researchers must first be approved before any data may be acquired (please see Appendix Table 3 for an image of the electronic application form). Moreover users are never permitted access to data containing the original codes provided by the National Statistical Institutes. Instead, only integrated data are provided, and these are only in the form of extracts, custom tailored to each researcher's needs. What this means is that there is no distribution of entire datasets by means of compact discs. Since each dataset is custom tailored “collecting” or “boot-legging” datasets is not only illegal, but effectively curtailed.

See Appendix Table 3

To request an extract, the researcher must first sign in by entering the registered password, then a series of selections are made by means of point-and-click menus. The researcher selects the country or countries, census years, samples, and variables as well as the form of metadata required for the statistics package to be used (SAS, SPSS, or STATA are supported). The IPUMS-International extract engine also makes it possible to select sub-populations, such as, say, females aged 15-19 in the workforce.

One of the most valuable enhancements of the database is the “SUBSAMPLE” feature. With SUBSAMPLE, the research may request any of 100 subsamples each of which is nationally representative and preserves any stratification of the larger sample from which it was drawn. This tool may be used to test procedures, economize resources, where the research does not require large samples, or estimate variances through the replicate method.

Once the selections are complete, there is an opportunity to review or revise before final submission of the request. Then, once submitted, the extract engine registers the request and places it in a data processing queue. When the extract is ready (usually in a matter of minutes), the researcher is notified by email that the data should be retrieved within 72 hours. A link is provided to a password-protected page for downloading the specific extract. Soon an SSL (Secure Sockets Layer) protocol will be implemented at the Minnesota Population Center. After SSL is in place, the data will be encrypted during transmission using a 128-bit encryption standard, matching the level used today by the banking and other industries where security and confidentiality is essential. The researcher may then securely download the file, decompress it and proceed with the

analysis using the supplied integrated metadata consisting of variable names and labels. The metadata are in ASCII format so that a researcher may readily adapt them for use by any statistical software.

Users and Uses. The IPUMS-International project offers bona fide researchers custom-tailored extracts at no charge via the Internet. During the first 33 months of operation, 766 applications were received, of which 39% were denied. The principal reason for rejecting an application is that the proposed research (as described by the applicant in the registration request—see Appendix A) does not seem to require access to the available microdata. In some cases, researchers request microdata for countries which are not presently integrated in the database (“hope springs eternal”). In others the proposed analysis requires information, such as certain environmental or economic variables, that is not present in the data. Then too, because of the anonymization methodology, fine-grained geographic identifiers are suppressed so requests requiring information about localities, villages, or even towns, must also be rejected. In each case, the reason for rejection is communicated to the researcher, so that a revised application may be re-submitted, if desired.

The following statistics are derived from applications for access of the first 469 approved users of the IPUMS-International database. Please note that incomplete applications are not included in this tally nor is any supplemental information which may have been requested from applicants in weighing a decision on whether to grant access or not. Approval is based solely on criteria of scientific feasibility (that census microdata are essential for the proposed research), including credentials of the researcher.

Who uses the data and what do they use it for? The succinct answer is university professors, students and policy researchers use the data to investigate economic, demographic and social issues in comparative perspective.

In a very brief period, IPUMS-International has become an indispensable component of social science infrastructure. Hundreds of projects by scholars in more than thirty-four countries are already underway. The United States accounts for the largest number of applicants (72%), followed by Canada (4%) (see Table 1). Switzerland, thanks to the presence of a large number of international organizations, ranks third (3%). Every continent is represented. Over 5% of researchers are working in Europe. Asian users, at less than 2% of the total, are under-represented at present, but this is because only 3 samples are included from the region, constituting less than 1/20th of the total person records in the database.

The application does not inquire as to country of origin, citizenship or identity. Nevertheless, it is apparent from names and project descriptions, that a considerable fraction of researchers at US and Canadian universities are nationals using the IPUMS-International database to study their country of origin, including not only Mexico, Colombia, Kenya and Vietnam but also France.

Countries of research interest. Research interest is limited to countries in the database at the time of the application. Researchers often express interest in countries for which no data are available, such as India, but these are not included in the following table. Brazil is tallied only if the country was specifically mentioned in the project description.

The most noteworthy point here is that 72% of approved projects indicated comparative research, involving more than one country. Percentages indicate the proportion of approved applicants expressing an intention to study a specific country, excluding applications before August 2002, when this information first began to be requested by means of a check box.

Use of the database to study Mexico and the United States stands out, particularly by researchers interested in studying Mexicans, regardless of whether they reside in their country of birth or in the United States. It is gratifying that for both France and Colombia, many researchers are using the IPUMS-International database rather than the national sources of census microdata. Thanks to the easy availability of the data and documentation, preliminary preparations are reduced from a matter of weeks or even months, to a day or two. All the information is on the website.

Country of residence	%	Country/ies of interest	%
USA	72	Brazil (since Sept. 2004)	4
Canada	4	China (since May 2003)	11
Switzerland	3	Colombia	13
Brazil, Colombia, Kenya (total)	8	France	12
France, Italy, Mexico, Spain, UK, Vietnam (total)	6	Kenya	12
China (includes Hong Kong, etc.)	1	Mexico	20
Australia, Germany	1	USA (excludes IPUMS-USA)	17
19 other countries (total)	5	Vietnam	11

Institutional affiliation and position (Table 2). Almost 90% of users are university based, and almost half the users are students. It is gratifying to see that researchers at national policy institutes are using the IPUMS-International harmonized microdata in preference to the privileged access that they often enjoy to data from their own country. National statistical agencies are registering with the idea of evaluating the site rather than doing research.

Institutional affiliation	%	Position	%
University	88	Student	48
Regional/International organization	8	Researcher	26
National policy institute	2	Professor	21
National statistical agency	2	Other	6

“Field” (academic discipline) is a “radio dial,” which means that applicants must select from among the options available. One-third of users are economists, followed by demographers at one-fourth. “Outcome” is inferred from the project description, which means that many users do not state what they expect to produce (57%). Most of the usage is for teaching (16%), followed by papers (10%), dissertations (9%) and finally

books (2%). A common, but somewhat surprising application, is to complement survey data (DHS, Employment, special one-of-a-kind surveys, etc.), to estimate population weights for the surveys.

Academic discipline	%	Expected outcome	%
Economics	37	Teaching, B.A./M.A. thesis	16
Demography	26	Paper, article, policy report	10
Sociology	13	PhD dissertation	9
Public policy	6	Book	2
History	5	Enhance DHS/other survey	6
Other	13	Other, Not mentioned	57

Funding. While there is no charge to draw extracts from the IPUMS-International extracts, researchers require funding, nevertheless. Funding agencies include CNRS (France), CONACYT (Mexico), National Institute of Aging (USA), CRF-VI (European Union), Fulbright (USA), SSHRC (Canada), CNPq (Brazil), IUT (France), World Bank, USAID, Andrew W. Mellon (USA), NSF (USA), NIH (USA), CAPES (Brazil), ESRC (UK), UNOSAT (Geneva), ANU (Australia), IDB (Washington DC), CIHR (Canada), CONICET (Argentina), etc.

Research topics. Applicants are required to submit a brief description of the proposed research. I have classified these, somewhat arbitrarily, into 26 categories (Table 4). They demonstrate the wide range of research uses for which census microdata may be used.

Migration	64	Marriage	12
Schooling	57	Aging	12
Gender	30	Equality/inequality	12
Data management/development	26	Mortality	12
Teaching	37	Development	10
Health	21	Statistics	9
Fertility	21	Sampling	9
Methods	17	Demography	7
Wages	17	Brain drain/gain	6
Urbanization	15	Religion	4
Family	15	Population projection	3
Children	13	Disability	3
Poverty	12	Vital statistics evaluation	2

Research topics include the living arrangements of the aged, female labor-force participation and educational attainment, regional inequality differentials, patterns of age hypergamy, international migration, relationship between divorce and family composition, between disease factors and education, and between marriage and socio-economic conditions. Most of these studies incorporate both cross-national and cross-temporal comparisons. For example, a National Academy of Sciences panel on “Transitions to Adulthood in Developing Countries” is using the data from Colombia, Kenya, Mexico, and Vietnam to analyze changing outcomes such as schooling, work, fertility, and marriage as a function of age, gender, and household characteristics. A scattering of studies propose to analyze various needs at the level of minor administrative districts for various institutions or professions, such as schools, teachers, clinics, health professionals, etc. While one might expect that these studies would be better served by access to 100% microdata, the 10% harmonized samples available from the IPUMS website make the results of such studies suggestive if not conclusive.

The following abridged and edited project description is a good example of a policy study which couples economic data from an official source with census microdata over four decades:

analyze the impact of public investment in [Country N] on a number of social and economic indicators over the last 40 years at the [major administrative district, MADs] level. There is evidence that despite high periods of overall growth in [Country N] very little economic convergence across [MADs] has occurred. This phenomenon has raised questions about the lack of ability (or willingness) of the central government to reduce disparities using national resources. This study tries to estimate the impact of different kinds of national investment and the role they have played over four decades of development in [Country N].

A quite different example is provided by a researcher who applied to the IPUMS-International database to design a new system for delivering harmonized census microdata. As far as we are aware the researcher has not yet implemented the system.

Conclusion. Now that the construction of anonymized microdata data samples is becoming an increasingly widespread practice, integration of census microdata is an obvious next step to enhance use. With the emergence of global standards for harmonizing census data and the massive power of ordinary desktop computers, the major challenge that remains is the actual construction of integrated census microdata samples. Thanks to the cooperation of some 50 official census agencies worldwide and with the financial support of the National Science Foundation and the National Institutes of Health, the IPUMS-International project is committed to integrating microdata for 150 censuses by 2010. If the IPUMS-International project is truly successful it will continue beyond the 2000 round of censuses, incorporating samples of participating countries for the 2010 censuses shortly after they become available. The number of users and uses may increase proportionately as well.

References.

Dale, A., Fieldhouse, E. and Holdsworth, C. (2000) *Analyzing census microdata*. Arnold: London.

- Esteve, Albert and Matthew Sobek. (2003). Challenges and Methods of Census Harmonization. *Historical Methods* 36: 66-79.
- McCaa, R. and Ruggles, S. (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, Miriam King, Deborah Levison, Robert McCaa, and Matthew Sobek. (2003). "IPUMS-International: An Overview". *Historical Methods*, 36: 60-65.
- United Nations Statistics Division. (1998). *Principles and recommendations for population and housing censuses*. Department of Economic and Social Affairs, New York.

Table 1. IPUMS-International Country Partners and Census Microdata, February 1, 2005

	Place	2000s	1990s	1980s	1970s	1960s
Key: »» = microdata for all census years entrusted to project						
Pending = agreement in principle, but official endorsement of MoU is pending						
Year = census conducted;						
Bold year = microdata survive; m = microcensus						
Phase I, 1999-2004						
»»	Brazil	2001	1991	1980	1970	1960
	China (only '82 included so far)	2000	1990	1982		1964
»»	Colombia		1993	1985	1973	1964
»»	France	1999	1990	1982	1975	1968, 62
»»	Kenya	1999	1989	1979	1969	
»»	Mexico ('80 not recovered yet)	2000	1990	1980	1970	1960
»»	United States	2000	1990	1980	1970	1960
»»	Vietnam		1999	1989	1979	
Phase II, 2004-2009						
Asia and the Pacific						
	Armenia	2001		1989	1979	
	Bangladesh	2001	1991	1981	1974	1961
»»	Cambodia		1998			1962
	Georgia	2002		1989	1979	
pending	Iran		1996	1986	1976	1966
	Iraq		1997	1987	1977	1967
	Israel		1995	1983	1972	1961, 67
»»	Malaysia	2001	1991	1980	1970	1960
»»	Mongolia	2000		1989	1979	
	Pakistan		1998	1981	1973	1961
»»	Palestinian Authority		1997			
»»	Philippines	2000	1990	1980	1970	1960
	Tajikistan	2000		1989	1979	
	Turkmenistan		1995	1989	1979	
Europe						
	Austria	2001	1991	1981	1971	1961
»»	Belarus		1999	1989		
	Bulgaria	2001	1992	1985	1975	1965
	Czech Republic	2001	1991	1980	1970	1961
	Germany (Ro and DR)	2001m	1991m	1987, 81	1970, 71	1961
	Greece	2001	1991	1981	1971?	1961
»»	Hungary	2001	1990	1980	1970	
	Ireland	2001	1991	1981	1971	1961
	Netherlands	2001m			1971	1960
pending	Poland	2001		1988	1978, 70	1960
	Portugal	2001	1991	1981	1970	1960
»»	Romania	2001	1992		1977	1965
pending	Russia (-1989 USSR)	2002	1994m	1989	1979	1970
	Slovenia	2001	1991	1981		
»»	Spain	2001	1991	1981	1970	1960
pending	Turkey	2000m	1990	1980, 85	1970, 75	1960, 65
	United Kingdom	2001	1991	1981	1971	1961

North America						
	Canada	2001	1991, 96	1981, 86	1971, 76	1961, 66
»»	Costa Rica	2000		1984	1973	1963
»»	El Salvador		1992		1971	1961
	Guatemala	2003	1994	1981	1973	1964
	Honduras	2000		1988	1974	1961
	Nicaragua		1995		1971	1963
»»	Panama	2000	1990	1980	1970	1960
South America and the Caribbean						
	Argentina	2001	1991	1980	1970	1960
»»	Bolivia	2001	1992		1976	
»»	Chile	2002	1992	1982	1970	1960
	Dominican Republic	2003	1993	1981	1970	1960
»»	Ecuador	2001	1990	1982	1974	1962
»»	Paraguay	2002	1992	1982	1972	1962
»»	Peru		1993	1981	1972	1961
»»	Puerto Rico	2000	1990	1980	1970	1960
»»	Venezuela	2001	1990	1981	1971	1961
Africa						
	Egypt		1996	1986, 81	1976	1964
	Madagascar		1993		1975	1966
	Malawi		1998	1987	1977	1966
»»	South Africa	2001	1996, 91	1985, 80	1970	1960
	Uganda	2000	1991	1980		1969
Datasets per Census Round (n)		45	53	38	36	18

Table 2. IPUMS-International Integrated Census Microdata Sample Characteristics			
120 million person records			
Source: www.ipums.org/international/sample_descriptions.html			
Country census	Sample %	No. of Person records	Additional details
Brazil 1960	5.0	3,001,000	Long-form, cluster sample
1970	5.0	4,954,000	Same
1980	5.0	5,870,000	Same
1990	5.0	8,523,000	Same
2000	6.0	10,136,000	Same
China 1982	0.1	1,003,000	Every thousandth household
Colombia 1964	2.0	350,000	Every fiftieth person
1972	10.0	1,989,000	Every tenth household
1985	10.0	2,643,000	Long-form, cluster sample
1993	10.0	3,247,000	Every tenth household
France 1962	5.0	2,321,000	Every twentieth household
1968	5.0	2,488,000	Same
1975	5.0	2,629,000	Same
1982	5.0	2,714,000	Same
1990	4.2	2,361,000	Every twenty-fourth household
Kenya 1989	5.0	1,074,000	Every twentieth household
1999	5.0	1,410,000	Same
Mexico 1960	1.5	503,000	Every 67th individual
1970	1.0	483,000	Every hundredth household
1990	10.0	8,028,000	Every tenth household
2000	10.6	10,099,000	Long-form, cluster sample
USA 1960	1.0	1,800,000	Stratified, random sample
1970	1.0	2,030,000	Same
1980	5.0	11,337,000	Same
1990	5.0	12,500,000	Stratified, cluster sample
2000	5.0	14,082,000	Same
Vietnam 1989	5.0	2,627,000	Long-form, cluster sample
1999	3.0	2,368,000	Same


[Home](#)
[Data](#)
[Documentation](#)
[Feedback](#)
[Search](#)

Project Description

[Principles](#)
[Progress Report](#)
[Release Dates](#)
[Revision History](#)

Data

[Apply for Access](#)
[Create an Extract](#)
[Download Extracts](#)
[Citation](#)

Documentation

[Samples](#)
[Variables](#)
[Source Materials](#)

Resources

[Microdata Inventory](#)
[Enumeration Forms](#)
[Microdata Handbook](#)
[International Partners](#)

Contact Us

IPUMS-International Data Extraction System

Application to Use Restricted Microdata

IPUMS-International microdata are available free of charge, but their use imposes responsibilities upon the user. To access the data from the Integrated Public Use Microdata Series-International site, a prospective user must first submit an electronic authorization form (this form) identifying the user by name, electronic address, and institution. The investigator must state the purpose of the proposed project and agree to abide by the regulations specified below. If multiple investigators are involved in a project, all must register separately. Once a project is approved, a message will be sent by email granting access to the system. The notification licenses the user to acquire microdata from Integrated Public Use Microdata Series International or other authorized distributors. No titles or other rights are conveyed to the user.

All information will be kept confidential.

All information on this form is required for registration.

Personal Information

First Name:

Last Name:

Employer/Institutional Affiliation:

Funded research, other than employer, if any:

Indicate name of granting institution, grant #, and year(s) of award, or state "None":

Does your institution have a Data Safety Monitoring Board, Office for Human Research Protections, or Professional Conduct Committee?

Yes

No

Address:

Street Address 1:

Street Address 2:

City, State/Province, Zip:

Country:

Phone Number(s): (include country and area codes)

Fax Number: (optional)

E-mail address:

Field:

Demography

Status:

Faculty

- | | |
|--------------------------------------|---|
| <input type="radio"/> Economics | <input type="radio"/> Academic researcher |
| <input type="radio"/> History | <input type="radio"/> Support staff |
| <input type="radio"/> Sociology | <input type="radio"/> Student |
| <input type="radio"/> Other academic | <input type="radio"/> Non-Academic Researcher |
| <input type="radio"/> Public Policy | |

Usage License

for Integrated Public Use Microdata Series International (IPUMS-International) and its partners

Please check all of the following boxes to indicate that you have read about the limitations of the IPUMS-International data and you agree to abide by the conditions of use. The purpose of this license is to specify the terms and conditions under which integrated microdata samples distributed by Integrated Public Use Microdata Series International of the University of Minnesota may be used.

Data must not be redistributed without authorization.

All data extracted from the IPUMS-International database are intended solely for the use of the licensee. Under IPUMS-International agreements with collaborating agencies, redistribution of the data to third parties is prohibited.

The microdata are intended only for scholarly research and educational purposes.

These microdata are provided for the exclusive purposes of teaching and scholarly research, and may not be used for any other purposes without explicit written approval.

Commercial use and redistribution of the microdata is strictly prohibited.

Users are prohibited from using microdata acquired from the Integrated Public Use Microdata Series International or other authorized distributors in the pursuit of any commercial or income-generating venture either privately, or otherwise.

Use of the microdata must follow strict rules of confidentiality.

Users will maintain the confidentiality of persons and households. Any attempt to ascertain the identity of persons or households from the microdata is prohibited. Alleging that a person or household has been identified in these data is also prohibited.

The microdata must always be safely secured.

Users will implement security measures to prevent unauthorized access to microdata acquired from Integrated Public Use Microdata Series International, its partners or authorized distributors.

Scholarly publications are permitted, and must be cited appropriately.

The publishing of research results based on IPUMS-International microdata is permitted in communications such as scholarly papers, journals and the like.

The authors of these communications are required to cite Integrated Public Use Microdata Series-International as the source of the microdata, and to indicate that the results and views expressed are those of the author. Users are asked to provide the IPUMS-International staff with a full citation for any publications resulting from their work with these data.

Any violation of this license agreement will result in disciplinary action, including possible loss of employment.

Violation of this agreement will lead to a revocation of this license, recall of all microdata acquired, a motion of censure to the relevant professional organization(s) and civil prosecution under the relevant national or international statutes, at the discretion of the Regents of the University of Minnesota and the national statistical agencies.

Description of Project Proposal:

Please provide a clear, substantial description of your research project or educational use for the data. This description will be used to evaluate your application.

Which country samples do you intend to use in your research?

- | | |
|-----------------------------------|--|
| <input type="checkbox"/> Brazil | <input type="checkbox"/> Kenya |
| <input type="checkbox"/> China | <input type="checkbox"/> Mexico |
| <input type="checkbox"/> Colombia | <input type="checkbox"/> United States |
| <input type="checkbox"/> France | <input type="checkbox"/> Vietnam |

The addition of data for what country(s) or region(s) would you consider most valuable for your future research?

Contingent upon acceptance of the application, your User Name will be set to the following email address: (Please make sure it is correct; change at the top of this form.)

Please enter your Preferred Password:

(at least 7 characters, using at least one alphabetic and one numeric character each)

Confirm Password:

Submit Registration Information