

Disseminating Anonymized, Integrated Census Microdata via the Internet: the IPUMS-International Project

Robert McCaa, Steven Ruggles and Matt Sobek

Minnesota Population Center

20th Population Census Conference, Ulaanbaatar, Mongolia – 19–21 June, 2002

Introduction. Census microdata samples are an invaluable resource for social science and policy research. Other 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. For much of the world, census microdata are either unavailable or restricted, and are therefore seldom used. Appendix 1 shows that statistical confidentiality is governed by national law in each of the fifty-two member-states of the International Monetary Fund’s General Data Dissemination System. Nevertheless, forty countries make census microdata available to researchers. The problem is that even where census microdata can be obtained, comparison across countries or time periods is challenging because of inconsistencies between datasets and inadequate documentation of comparability problems. Because of this, comparative chronological or international research based on pooled census samples is rarely attempted.

The IPUMS-International project is reducing the barriers to international research by anonymizing census microdata, converting it into a uniform format, providing comprehensive documentation, and making the data freely available through a web-based access system to bona-fide researchers who sign a legally binding non-disclosure agreement. The first release, now available at the project website (www.ipums.org/international), consists of 21 samples for six countries: Colombia (1964-1993), France (1962-1990), Kenya (1989-99), Mexico (1960-2000), Vietnam (1989-99), and the USA (1960-1990). A second release, planned for 2004, will add samples for Brazil, China, Hungary, Spain, and Ghana. Now, the project is beginning 5-year regional initiatives with the first in Latin America, where all nations, with the exceptions of Cuba and Uruguay, have agreed to join the project and adopt uniform standards of statistical confidentiality, integration and distribution.

This paper, first, summarizes the principles of the project: anonymization and statistical confidentiality, harmonization, and distribution of the data. Second, the basic work plan is discussed, with some reference to the project for the Latin America region. On behalf of the director of the Minnesota Population Center, Dr. Steven Ruggles, the presenter thanks the organizers for the opportunity to describe the IPUMS-International project. He welcomes the opportunity to discuss possible partnerships with national statistical agencies associated with ANCSDAAP.

IPUMS-USA. The Integrated Public Use Microdata Series (IPUMS-USA) is partly responsible for the widespread use of census microdata by demographers studying the United States. IPUMS-USA, developed by Steven Ruggles, Matthew Sobek, and others at the Minnesota Population Center, makes anonymized census microdata samples freely available to scholars in harmonized format with comprehensive documentation 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. Over 6,000 researchers have

registered to use the IPUMS data extraction system. The user base continues to expand rapidly, with approximately 2,500 new registered users during the past year alone. We are now distributing about 140 gigabytes of data per month, or an average of 190 megabytes per hour, twenty-four hours a day. We have prepared approximately 60,000 custom extracts of IPUMS data since May 1996 and are now processing approximately 2,800 data extract requests per month. This massive data distribution is beginning to bear fruit. Although the IPUMS has been available for only six years, at this writing our bibliography lists twenty-six books, seventy-one dissertations, 207 published research articles, and hundreds of working papers, conference presentations, and research reports (<http://ipums.org/usa/research.html>).

In Canada as well as the United States census microdata have been available to researchers for almost forty years and have become an indispensable component of social science infrastructure. For example, census microdata were the data source for nineteen of the fifty-one 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.

IPUMS-International. In 1998 the Minnesota Population Center proposed to extend the IPUMS paradigm to the censuses of Colombia (R01HD37508). This pilot project, a collaboration with the Colombian National Statistical Office (DANE), was designed to demonstrate the feasibility of creating public use microdata for Latin America. Shortly after we proposed the Colombia project, the National Science Foundation announced a special program for “Enhancing Infrastructure for the Social and Behavioral Sciences” that offered one-time funding for major new data improvement initiatives. We proposed a large-scale international project with two major components (SBR9907416). The first step was to identify and preserve surviving machine-readable census microdata from around the world for the period 1960 to 2000. The second step was to select seven countries with broad geographical distribution and to clean, harmonize, document, and disseminate microdata for those countries using the same principles and methods that underlie the original IPUMS-USA database.

These two international projects, collectively known as IPUMS-International, have been an unqualified success, thanks to the cooperation of National Statistical Agencies, the United Nations Statistics Division and other agencies. Both projects are now in their third year and are well ahead of schedule. We have created a comprehensive inventory of known microdata, much of which is described in the award-winning book, *Handbook of International Historical Microdata* (Hall, McCaa, and Thorvaldsen 2000), and we have preserved microdata from over one hundred censuses. In May 2002, we released our first group of harmonized census microdata samples for Colombia, France, Kenya, Mexico, the United States, and Vietnam (<http://ipums.org/international>). In 2004, we plan to release a second group of harmonized samples for Brazil, China, Ghana, Hungary, and Spain.

Our first release of international census microdata samples has been available for only a few weeks, and publicity for the samples has been mainly word-of-mouth. Nevertheless, the reaction of scholars to the new data has been so enthusiastic that we anticipate IPUMS-International will soon rival the usage statistics of IPUMS-USA. We have already received dozens of applications for

access to the data from scholars in the United States, Panama, Norway, Kenya, Hungary, Switzerland, and Canada. In addition to university-based researchers, the user list includes representatives of four national statistical offices 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 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. A National Academy of Sciences panel on “Transitions to Adulthood in Developing Countries” is using the data from Colombia, Kenya, Mexico, and Vietnam. The goal of this panel is to analyze changing outcomes such as schooling, work, fertility, and marriage as a function of age, gender, and household characteristics.

Despite the important contribution of IPUMS-International, it has limitations. Funding was provided to create samples for just a scattering of countries around the globe. Moreover, those countries are so different from one another—with respect to both their census definitions and procedures and their social norms and behavior—that cross-national comparisons are difficult. To fully capitalize on the potential of international census microdata, a more focused regional approach is needed. The next stage of the project is to develop sub-continent or continental integration initiatives, such as the Latin America project.

Confidentiality protection. The protection of respondent confidentiality is of paramount importance. We use two strategies for safeguarding the confidentiality of microdata: confidentiality/licensing agreements (see Appendix 2: “Letter of Understanding”) and statistical disclosure protections. Used in combination, these approaches minimize the potential risk of disclosure without seriously compromising scientific use of the data.

Some thirty countries have now agreed to the IPUMS-International framework of safeguards for distributing microdata. We disseminate microdata only under strict confidentiality controls approved by each national statistical office. Before data are released, individual researchers must submit 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 the information on to the relevant national statistical agencies.

In addition, researchers must propose a research project that demonstrates a scientific need for the microdata. Each application for access is evaluated by senior staff. Once an application is approved, the user password is activated, allowing controlled access to data. Penalties for violating the license include revocation of the license, recall of all microdata acquired, filing a motion of censure to the appropriate professional organizations, and civil prosecution under the relevant

national or international statutes. Employees of the Minnesota Population Center who work with the census microdata also sign agreements to respect the confidentiality of the data.

The National Institutes of Health of the United States requires the filing of a Human Subject Protection statement to accompany the grant application (see appendix 3). Every funded research project must abide by these regulations. Violation is a serious crime which may be punished by imprisonment or fine or both.

Technical safeguards supplement these institutional controls. 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, we use a variety of other procedures 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.

Harmonization. International census samples employ differing concepts and numeric classification systems, and reconciliation of these is a major part of the project. Variable design often influences the analytical strategies adopted by researchers, and we have therefore developed our plans with care.

United Nations organizations have twice sponsored large-scale projects for regional harmonization of census microdata. The first was the OMUECE project sponsored by the Centro Latinoamericano de Demografía (CELADE). CELADE standardized versions of twenty-nine Latin American censuses taken between 1960 and 1976 (McCaa and Jaspers 2000). The second project was undertaken by the United Nations Population Activities Unit (PAU) in Geneva (Botev 2000). This project, which is still ongoing, is a standardization of microdata from the 1990 and 2000 rounds of censuses of fifteen European and North American countries. These two initiatives have provided IPUMS-International with valuable information. They have allowed us to take advantage of the investments already made by the United Nations and to learn from the experience of earlier attempts at international census harmonization.

The IPUMS-International design strategy is more ambitious than that of either CELADE or PAU. Unlike CELADE, we retain all the detail provided in the original samples. Unlike PAU, we provide a truly integrated database, in which identical categories in different census samples always receive

identical codes. We employ several strategies 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 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.

To take the simplest example, the classification scheme for marital status illustrates this point. Under the IPUMS-International design, 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.

Dissemination. Data access is an integral component of the project; effective dissemination is essential if the data are to be widely used. A complete set of documentation and integrated census microdata is provided to the National Statistical Agency, which may distribute the data as it wishes. For researchers, both data and the documentation are distributed through an integrated web-based data access system at no charge. We have been working on methods of electronic dissemination for social science data and documentation for almost a decade. We have already developed the most powerful web-based data extraction system available for access to large microdata files. The IPUMS-USA data access system pioneered web-based dissemination of large-scale data, and it has served as a model for many other social science data dissemination efforts. This research experience provides the foundation for our current efforts to improve data sharing technology.

IPUMS-International is now developing second-generation data dissemination software. The new data access system provide advanced tools for navigating documentation, defining datasets, constructing customized variables, and adding contextual information. A preliminary version of this system is already operating for the first set of IPUMS-International data. This secure data extraction system allows users to merge datasets, subset populations, and select variables. Documentation browsing functions are built into the data extraction tool so that users have easy access to comprehensive documentation as they design their analyses.

Work plan. The project is a partnership between the Minnesota Population Center, the National Statistical Agencies and national experts. Our data dissemination agreements and license fees provide not only for dissemination rights, but also for the supply of ancillary materials (such as codebooks and technical publications) and technical support by the staff of these agencies. As needed, the project also supplements this pool of knowledgeable specialists with other experts drawn from across the region. They answer questions on census enumeration procedures and post-enumeration data processing, the methodology employed to create existing samples, and specific

integration problems (such as the details of economic, education, housing, and geographic variables for particular countries).

The tasks to be accomplished may be summarized as follows:

1. Assemble complete documentation (enumeration forms, enumerator instructions, codebooks, etc) and microdata.
2. Clean raw data files (e.g., identify and correct data format problems; carry out internal consistency checks; identify coverage problems through comparison with published statistics).
3. Draw samples from internal census files.
4. Impose confidentiality protections (e.g., top-codes, geographic swapping, category blurring, and randomization of household sequence within geographic units).
5. Recode variables into the IPUMS-International harmonized coding system to permit analysis across countries and time periods; develop and apply new harmonized coding designs, as needed.
6. Allocate missing and inconsistent data values through probabilistic and logical editing procedures.
7. Create a set of consistent constructed variables describing household composition, family interrelationships and socioeconomic status.
8. Develop harmonized English-language documentation (e.g., census enumeration procedures and instructions; post-enumeration processing; sample designs; variable-level documentation on census questions, universe definitions, variable category availability, and frequency distributions; definitions of households, dwellings, group quarters and other enumeration units; and comparability issues across census years and countries).
9. Convert all documentation to the Data Documentation Initiative (DDI) international metadata standard.
10. Disseminate data from IPUMS International web-site. Provide copies on CD of the nationally integrated microdata to the National Statistical Agencies.

IPUMS-Latin America. Table 1 describes the censuses to be incorporated in the Latin America database. The table includes the two countries—Bolivia, and Uruguay—that have not yet signed the agreement, as well as the Commonwealth of Puerto Rico, whose data are in the public domain. The left panel reports the percentage of cases that survive for each census. For twenty-seven of the censuses taken from the 1970s to the 1990s, we have complete data. For the thirty-five remaining censuses, only samples of microdata survive, with densities ranging from 1 to 25 percent. All but four of these samples relate to the 1960 and 1970 census rounds. The decade of the 1960s is the weakest part of the data series; only sample data survive in machine-readable form, and most of those samples were taken at the individual level and are non-hierarchical.

Conclusion. Now that the population census has become a global phenomenon, and the construction of anonymized microdata data samples an increasingly widespread practice, harmonization of census microdata is an obvious next step to enhancing use. With the emergence of global standards of statistical confidentiality and the massive power of ordinary desktop computers, the only remaining obstacle is the integration of anonymized census microdata samples. If the

experiences of Canada and the United States are reliable guides, an explosion in scholarly and policy research is likely to ensue.

References.

- Botev, Nikolai. 2000. PAU Census Microdata Samples Project. In *Handbook of International Historical Microdata for Population Research*, edited by Patricia Kelly Hall, Robert McCaa and Gunnar Thorvaldsen. Minneapolis: Minnesota Population Center, pp. 303-17.
- McCaa, Robert, and Dirk J. Jaspers-Faijer. 2000. The Standardized Census Sample Operation (OMUECE) of Latin America, 1959-1982 [1995]: a Project of the Latin American Demographic Center (CELADE). In *Handbook of International Historical Microdata for Population Research*, edited by Patricia Kelly Hall, Robert McCaa, and Gunnar Thorvaldsen. Minneapolis: Minnesota Population Center, pp. 287-302.
- 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. Data User's Perspective on Confidentiality. *Of Significance . . . Journal of the Association of Public Data Users* 2:1-5.
- Ruggles, Steven, and Matthew Sobek, et. al. 1997. *Integrated Public Use Microdata Series: Version 2.0*. Minneapolis: Historical Census Projects, University of Minnesota.

Appendix 1. Statistical Confidentiality and Census Microdata Dissemination Practices			
Synthesis of Confidentiality Provisions, 52 member-states:			
Country	Law	International Monetary Fund's General Data Dissemination System	Census Microdata
Argentina	1968	Individual reports and/or data may not be communicated to third parties or used or disseminated in such a way as to make it possible to <u>identify the reporting person or entity</u> .	CELADE
Australia	1905	The Census Act protects the confidentiality of persons and organisations by requiring that information not be published in a manner likely to enable the identification of a particular person or organisation. Notwithstanding this, the CSA provides for the Minister to make determinations providing for the release of certain classes of information which would not otherwise be permitted to be released under the Act; except that <u>personal or domestic information may not be disclosed under the provisions of a determination in a manner that is likely to enable the identification of a person (emphasis ours)</u> .	Australian National University
Austria	2000	Strict provisions on statistical confidentiality are contained in the Federal Statistics Act. The field on protection of personal data is covered by the Data Protection Act.	IPUMS ⁱ
Bangladesh		There are no regulations enforcing confidentiality of reporting, but <u>strict confidentiality is maintained in practice</u> .	
Belgium	1994	According to the rules of the Official Statistics Act..., the <u>confidentiality of individual responses is protected</u> .	ECE/PAU
Brazil	1999	Decree 74.084 of May 20, 1974... and Decree 3.272 of December 3, 1999...provide <u>assurances of confidentiality of individual responses</u> so that the data can be used only for statistical purposes.	CELADE IPUMS ⁱ
Canada	1985	[Under the Statistics Act of 1985.] Statistics Canada cannot publish, or otherwise make available to any individual or organization, statistics that would <u>enable the identification of data for any individual person or entity</u> .	ECE/PAU
Chile	1970	Law No. 17-374 and its Regulations All individuals and legal entities are required to provide any information requested by the INE, which in turn is required to <u>maintain strict confidentiality and is prohibited from explicitly referring directly or indirectly in its publications to individuals or legal entities</u> .	CELADE IPUMS ⁱ
Colombia	1960	Article 75 of Decree 1633 of 1960...establishes the <u>principles of confidentiality and discretion; thereby forbidding communication of data by name or individually</u> .	CELADE IPUMS ⁱ
Croatia	1994	Under Law N.N. 52/94, the CBS cannot publish, or otherwise make available to any individual or organization, statistics that would <u>enable the identification of data for any individual person or entity</u> .	
Czech Republic	2000	The State Statistical Service Act No. 89/1995 Coll. Which came into force on June 15, 1995 and was amended by Act No. 220/2000 Coll. And Act No. 411/2000 Coll. ... <u>Protection of individual data represents an important section of this Act</u> .	ECE/PAU
Denmark		According to the 'Public Authorities' Registers Act', <u>data attributable to identifiable individuals (or enterprises) shall not be passed on</u> .	
Ecuador	1976	The Official Registry Law No. 82 establishes the principles of confidentiality and discretion, thereby <u>forbidding disclosure of information for any individual person or private entity</u> .	CELADE IPUMS ⁱ
El Salvador	1955	...data compiled by the DIGESTYC are <u>confidential and may be used solely for statistical purposes</u> .	CELADE IPUMS ⁱ
Estonia	1997	The SOE may transmit or disseminate collected data only in a form which <u>precludes the possibility of direct or indirect identification of the respondents</u> .	ECE/PAU
Finland	1994	Under the terms of Act 62/1994, Statistics Finland cannot publish, or otherwise make available to any individual or organization, statistics that would <u>enable the identification of data for any individual person or entity</u> .	ECE/PAU
France	1978	INSEE cannot publish, or otherwise make available to any individual or	IPUMS ⁱ

		organization, statistics that would <u>enable the identification of data for any individual person or entity.</u>	
Germany	1987	[no specific statement on confidentiality.] (Collection and current updating of population data are regulated by the Law on the Statistics of Population Movement and Adjustment of the Population State dated March 14, 1980 in conjunction with the Law on Statistics for Federal Purposes of 1987.)	German Research Institute data enclaves
Hong Kong	1993	The 1978 Ordinance updated in 1993 stipulates that: ... (2) Only aggregate information will be published such that information relating to any particular individual or undertaking will be kept <u>strictly confidential and will not be divulged to other parties.</u>	EWC
Hungary	1993	The 1993 Law on Statistics of Hungary (XLVI/1993) and the 1992 Law on Protection of Personal Data and the Disclosure of Data of Public Interest (Law LXIII/1992) ... (4) All statistics collected and published by the HCSO are governed by the confidentiality provisions which specify that the HCSO cannot publish, or otherwise make available to any individual or organization, statistics that would <u>enable the identification of data for any individual person or entity.</u>	ECE/PAU IPUMS ¹
Iceland	2000	<u>Individual data are kept strictly confidential and care is taken that the data released cannot be traced directly or indirectly to an individual entity.</u> Researchers may be given access to information on individuals with the permission of the Data Protection Authority under strict rules and conditions.	
India	1948	Data relating to individuals have to be kept confidential.	Institute of Economic Growth
Indonesia	1997	The BPS (Law 16, 1997) cannot publish, or otherwise make available to any individual or organization, statistics that would <u>enable the identification of data for any individual or entity.</u>	EWC
Ireland	1983	The Statistics Act of 1993 ... sets stringent confidentiality standards: the information collected may be used only for statistical purposes, and <u>no information that could be related to an identifiable person or undertaking may be released.</u>	
Israel	1978	The Law on Statistics (1972 as amended in lawbook 908, 1978): ... (3) Stipulates that the CBS cannot publish, or otherwise make available to any individual or organization, <u>statistics that would enable the identification of data for any individual person or entity.</u>	ECE/PAU
Italy	1989	The Law on the National Statistical System (Legislative Decree n. 322, September 6, 1989) which is consistent with the U.N. Fundamental Principles of Official Statistics ... establishes: ... Strict confidentiality rules for data included in the National Statistical Program, approved yearly by Decree of the President of the Council of Ministers (D.P.C.M.) (<u>Dissemination occurs only in an aggregate form and in a manner by which it is not possible to identify data for any individual person or entity.</u>)	ECE/PAU IPUMS ^{1,*}
Japan	1999	Law to Establish the Ministry of Public Management, Home Affairs, Posts and Telecommunications (MPHPT) of July 16, 1999, and the Cabinet Order on the Organization of the MPHPT. ... - [no specific confidentiality statement on GDDS web-site.]	
Korea	1993	The Statistics Act of 1993 ... sets stringent confidentiality standards: the information collected may be used only for statistical purposes, and <u>no information that could be related to an identifiable person or undertaking may be released.</u>	EWC
Latvia	1997	The Law on State Statistics adopted on November 6, 1997 ... provides that the CSB cannot publish, or otherwise make available to any individual or organization, statistics that would <u>enable the identification of data for any individual person or entity.</u>	ECE/PAU
Lithuania	1999	Under the Law on Statistics (1999, No. VIII-1511) ... Statistics Lithuania cannot publish, or otherwise make available to any individual or organization, statistics that would <u>enable the identification of data for any individual person or entity.</u>	ECE/PAU
Malaysia	1989	Under the terms of the Statistics Act, 1965 (Revised 1989), DOSM: (2) Cannot	EWC

		publish, or otherwise make available to any individual or organization, statistics that would <u>enable the identification of data for any individual person or entity.</u>	
Mexico		All data provided by individuals or obtained from administrative or civil registers are treated with <u>strict confidentiality and discretion, and in no case may they be communicated by name or individually (Article 38).</u>	CELADE IPUMS ⁱ
Netherlands	1996	'Data gathered on the basis of this law will not be disclosed in such a form that returns and information about an individual person, company, or institutions can be deduced, unless the individual, the head of the company, or the governing board of the institution have no objection to such disclosure.'	
Norway	1989	Statistics Norway is <u>prohibited to publish or disclose data from which information about individual persons or firms can be derived.</u> (Researchers may be given access to such information under strict rules and conditions. Guidelines provided by the Norwegian Data Inspectorate form the framework for internal management of data security.)	ECE/PAU; Statistics Norway vetting of researchers
Peru	1990	INEI's Organization and Functions Law (Legislative Decree No. 604) of May 3, 1990 ... establishes the technical autonomy of INEI, details the norms concerning compilation of the data, and stipulates that information provided to the Peruvian statistical system is <u>confidential and cannot be disclosed individually,</u> even under an administrative or judicial order, and requires that the organization publish the data on population.	CELADE IPUMS ⁱ
Philippines	1987	The ... Commonwealth Act No. 591 (August 19, 1940), Executive Order No. 121 (January 30, 1987), and Batas Pambansa Blg. 72 (June 11, 1980). ... Section 4 provides that data furnished to NSO will be kept strictly confidential and shall not be used as evidence in court for purposes of taxation, regulation or investigation; nor shall such data or information be <u>divulged to any person except in the form of summaries or statistical tables in which no reference to an individual, corporation, association, partnership, institution or business enterprise shall appear.</u>	EWC
Poland	1995	Under the Law on Official Statistics, which was passed on 29 June 1995 (Dz. U. Nr. 88) ... the CSO cannot publish, or otherwise make available to any individual or organization, statistics that would allow the <u>identification of data of any individual person or entity.</u>	ECE/PAU
Portugal	1989	The National Law on Statistics (Law 6/1989 of April 15, 1989), ... establishes the principle of the technical independence of the INE, as well as the principle of confidentiality under which <u>no individual information about people can be disseminated.</u>	
Singapore	1991	The Statistics Act, Revised Edition, 1991 ... specifies that the disseminating agencies cannot publish, or otherwise make available to any individual or organization, statistics that would <u>enable the identification of data for any individual person or entity without prior consent.</u>	
Slovak Republic	1992	All statistical information collected, processed and released by SO SR is regulated by the Law on State Statistics (Law of SNC No. 322/92 Digest, in wording of latter regulations). This Law: ... - Specifies that individual responses to statistical surveys <u>cannot be used for other than statistical purposes without the permission</u> of the legal or physical person in question.	
Slovenia	1995	The Law on National Statistics ... (UrL RS No. 45/95) ... Emphasizes the importance of data confidentiality and stipulates that the Statistical Office cannot publish, or otherwise make available to any organization or individual, statistics that would <u>enable the identification of data for any individual person or entity.</u>	
South Africa	1999	The Statistics Act, 1999 (Act No. 66 of 1999) ... - Stipulates that Stats SA cannot publish, or otherwise make available to any individual or organization, statistics that would <u>enable the identification of data for any individual person or entity.</u>	ACAP
Spain	1996	Statistical Law No. 12/1989 ... and Law No. 13/1996: ... INE cannot publish, or make otherwise available, individual data or statistics that would <u>enable the identification of data for any individual person or entity.</u> (Article 13)	ECE/PAU IPUMS ⁱ
Sri Lanka	1981	The DCS produces and disseminates data under the Statistical Ordinance and	EWC

		Census Ordinance (1981) ... Confidentiality of reporters is guaranteed under the 1981 Ordinance which states '...no publication ... <u>shall disclose or facilitate the identification of any particulars as being particulars relating to any individual person</u> '.	
Sweden	1992	Data protection is ensured by prescriptions in the Data Act of 1973 (1973:289) and the Secrecy Act of 1980 (1980:100).	ECE/PAU
Switzerland	1992	The Federal Law on Data Protection (06/19/92) specifies that the Swiss Federal Statistical Office cannot publish, or otherwise make available to any individual or organization, statistics that would <u>enable the identification of data for any individual person or entity</u> .	ECE/PAU
Thailand		[No statement on confidentiality provided.]	EWC
Turkey	1989	The 1962 Statistical Law, as well as the 1984 Decree 219 and 1989 Decree 357: ... Data may be collected only for statistical purposes and confidentiality is assured. ... (3) The <u>confidentiality of individual responses is guaranteed</u> .	ECE/PAU
Uganda	1998	The Uganda Bureau of Statistics Act, 1998 ... Article 19 <u>ensures confidentiality of reported data</u> and Article 29 provides for substantial penalties to employees of the Bureau who violate the confidentiality provisions.	
United Kingdom		The Registrar General is required to compile and publish statistics on the number and condition of the population (1920 Census Act). Births and deaths from the National Registration System are subject to specific <u>statutory confidentiality constraints</u> , in addition to the general confidentiality policy of the ONS.	ECE/PAU CMCCSR IPUMS ¹
United States	1954	'No individual-level input data are released.' [Title 13 United States Code Section 9 prohibits 'any publication whereby the data furnished by any particular establishment or individual under this title can be identified'.]	IPUMS-USA ECE/PAU
Venezuela	1999	Law on National Statistics and Censuses of November 27, 1944 ... Article 10: 'The Ministry of Development may officially order aggregate or average data, or statistical series, but in no way and under no pretext may it order or authorize the <u>disclosure of individual data or the dispatch of single copies</u> ... related to a given individual or legal entity or to a given family or group of families.'	CELADE IPUMS ¹
Repositories of anonymized census microdata samples for scientific research			
Acronym	Institution and Dissemination Policy		
ACAP	African Census Analysis Project, Philadelphia USA. Permission of ACAP director.		
CELADE	Centro Latino Americano de Demografía, Santiago Chile. Application to National Statistical Agency.		
ECE/PAU	ECE Population Affairs Unit, Geneva Switzerland. Written application to Population Affairs Unit.		
EWC	East-West Center, Honolulu USA. Restricted to institution use only.		
ICPSR	Inter-University Consortium for Political and Social Research, Ann Arbor USA. Accessible by member universities.		
IPUMS¹	Integrated Public Use Microdata Series International, Minneapolis USA. Electronic application.		
CMCCSR	Cathie Marsh Center for Census and Survey Research, Manchester UK. Written application to CMCCSR.		
Note:	* = under negotiation.		
Sources:	For confidentiality provisions: International Monetary Fund (2001); microdata availability: Kelly Hall, et. al. (2000).		

Appendix 2. Letter of Understanding **Integrated Public Use Microdata Series International** and [National Statistical Agency of X]

Purpose. The purpose of this letter is to specify the terms and conditions under which metadata and microdata provided by the [National Statistical Agency of X] shall be distributed by **Integrated Public Use Microdata Series International** of the University of Minnesota.

1. Ownership. The [National Statistical Agency of X] is the owner and licensee of the intellectual property rights (including copyright) in the metadata and microdata supplied to the University of Minnesota to be distributed by **Integrated Public Use Microdata Series International**.

2. Use. These data are provided for the exclusive purposes of teaching, academic research and publishing, and may not be used for any other purposes without the explicit written approval, in advance, of the [National Statistical Agency of X].

3. Authorization. To access or obtain copies of integrated microdata of [X] from **Integrated Public Use Microdata Series International**, a prospective user must first submit an electronic authorization form identifying the user (i.e., principal investigator) by name, electronic address, and institution. The principal investigator must state the purpose of the proposed project and agree to abide by the regulations contained herein. Once a project is approved, a password will be issued and data may be acquired from servers or other electronic dissemination media maintained by **Integrated Public Use Microdata Series International**, the [National Statistical Agency of X], or other authorized distributors. Once approved, the user is licensed to acquire integrated metadata and microdata of [X] from **Integrated Public Use Microdata Series International** or other authorized distributors. No titles or other rights are conveyed to the user.

4. Restriction. Users are prohibited from using data 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.

5. Confidentiality. Users will maintain the absolute confidentiality of persons and households. Any attempt to ascertain the identity of a person, family, household, dwelling, organization, business or other entity

from the microdata is strictly prohibited. Alleging that a person or any other entity has been identified in these data is also prohibited.

6. Security. Users will implement security measures to prevent unauthorized access to microdata acquired from **Integrated Public Use Microdata Series International** or its partners.

7. Publication. The publishing of data and analysis resulting from research using metadata or microdata of [X] is permitted in communications such as scholarly papers, journals and the like. The authors of these communications are required to cite [**National Statistical Agency of X**] and **Integrated Public Use Microdata Series International** as the sources of the data of [X], and to indicate that the results and views expressed are those of the author/user.

8. Sharing. **Integrated Public Use Microdata Series International** will provide electronic copies to the [**National Statistical Agency of X**] of documentation and data related to its integrated microdata as well as timely reports of authorized users.

9. Violations. Violation of this agreement may lead to professional censure and/or civil prosecution.

10. Jurisdiction. Disagreements which may arise shall be settled by means of conciliation, transaction and friendly composition. Should a settlement by these means prove impossible, a Tribunal of Settlement shall be convened which will rule upon the matter under law. This Tribunal shall be composed of an (1) arbitrator, which shall be elected by lot from the list of Arbitrators of the Chamber of Commerce of Paris. This agreement shall be governed by, and construed in accordance with, generally accepted principles of International Law.

Date: _____

Signed: _____

Regents of the University of Minnesota

By: Kevin J. McKoskey, Sponsored Projects Administration

Date: _____

Signed: _____

Rev. Oct. 23, 2001

Appendix 3. Protection of Human Subjects (form submitted for IPUMS-Latin America project; format as required by the National Institutes of Health)

1. Risks to the subjects

Human Subjects Involvement and Characteristics. The study population consists of systematic samples of individuals within their households, who were enumerated in the national censuses that seventeen Latin American countries and the Commonwealth of Puerto Rico conducted between 1960 and 2003. The sample populations are representative with respect to the gender, age range, health status, and racial and ethnic composition of each country. The total number of cases in the database will consist of approximately 135 million records for individuals.

Sources of Materials. The project will make use of complete count census data from Latin American countries to draw samples of households and individuals. It will also use existing census microdata samples from these nations, when only sample data survive. The data were archived by each nation in the collection of the United Nations-sponsored Centro Latinoamericano y Caribeño de Demografía (CELADE). Samples from censuses conducted between 2000 and 2003 will be drawn by the national statistical agencies of collaborating nations.

Dissemination agreements have been negotiated with and signed by the national statistical agency of each participating country. These agreements provide for a license for dissemination of the census microdata by the Minnesota Population Center and other authorized distributors.

Potential Risks. Each national statistical office will deliver files to us that have already been anonymized. The names, addresses, and other potentially identifying information will be stripped off before the data arrive in Minnesota. While the data files will not include individual names or addresses, they may include sufficient geographic and subject detail to make identification of respondents a theoretical possibility. The potential risks to subjects from disclosure of census characteristics could include legal liability, risk to employment, or embarrassment.

2. Adequacy of protection against risks

Recruitment and Informed Consent. Informed consent is not applicable to national Censuses; in every country, residents are legally required to respond to censuses.

Protection Against Risk. Protection of respondent confidentiality is one of the highest priorities of the project. Each nation has a set of standards to ensure confidentiality, and these standards vary slightly from country to country. Under the signed dissemination agreements negotiated with each country, the Minnesota Population Center is legally bound to respect the standards set by each country, and to limit the variables and variable codes in the dataset as specified by the corresponding national statistical agency.

As noted, the national statistical offices and CELADE will deliver files to us that have been anonymized by stripping off names, addresses, and low-level geographic information. The Minnesota Population Center will take additional steps to ensure respondent confidentiality. As discussed in the section on confidentiality, we will take the following actions: randomizing the sequencing of records so that detailed geography cannot be inferred from position in the file; swapping an undisclosed fraction of records from one administrative district to another to make

positive identification of individuals impossible; combining codes that reveal sensitive characteristics or identify very small population subgroups (such as small ethnic categories); imposing bottom- and top-codes and rounding continuous variables (such as income). Employees of the Minnesota Population Center who work with the microdata sign agreements to respect respondent confidentiality. The effectiveness of these protections is likely to be great, based on the safety record for public use census microdata. Over the past four decades, there has not been a single verified breach of confidentiality for such data (Ruggles 2000).

In addition to these technical safeguards, we have a number of legal safeguards in place. As noted, we disseminate microdata 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 (<http://www.ipums.org/cgi-bin/ipumsi/ipumsireg.cgi>). To gain access to the data, researchers must agree to maintain the confidentiality all persons, households, and other entities. Any attempt to ascertain the identity of persons or households is prohibited, as is alleging that a person or household has been identified. Applicants agree to implement security measures to prevent unauthorized access to the data, and must not redistribute the data to third parties. The licensing agreement further specifies that the microdata must be used exclusively for scholarly research and education, and may not be used for any commercial or income-generating venture. Any publications based upon the data must be reported to the Minnesota Population Center, which will pass the information on to the pertinent national statistical agencies.

Potential researchers must propose a research project that demonstrates a scientific need for the microdata, and their proposal is evaluated by our senior staff. Once an application is approved, the user password is activated, allowing controlled access to the 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.

3. Potential benefits of the proposed research and importance of the knowledge to be gained

The potential benefits of the proposed database are described in this proposal. For example, increased understanding of such issues as the causes and correlates of fertility decline, population aging, and international migration from Latin America to the United States has potential benefit for all members of Latin American society, U.S. citizens, and social scientists and policymakers worldwide.

Table 1. IPUMS-Latin America: Density and Estimated Sample Sizes, by Country and Decade of Census

	Density of Source Microdata (%)					Person Records in Sample (000s)				
	1960s	1970s	1980s	1990s	2000s	1960s	1970s	1980s	1990s	2000s
Argentina	3	2	2	100	100	500	469	559	3,262	3,700
Bolivia	.	100	.	100	100	.	461	.	642	830
Brazil	25	25	25	12	10	7,028	9,252	11,752	14,205	17,000
Chile	1	5	100	100	100	88	443	1,133	1,335	1,520
Colombia	2	100	100	100	100	350	1,989	2,643	3,275	4,000
Costa Rica	6	100	100	.	100	82	187	242	.	360
Dominican Republic	7	7	8	100	100	203	272	476	761	840
Ecuador	3	17	100	100	100	136	924	835	965	1,260
El Salvador	1	5	.	100	100	26	176	.	512	630
Guatemala	5	5	5	100	100	210	290	302	833	1,270
Honduras	1	10	100	.	100	19	278	425	.	610
Mexico	1.5	1	n.a.	100	100	503	483	.	8,028	10,100
Nicaragua	n.a.	10	.	100	.	.	189	.	436	.
Panama	5	20	100	100	100	54	286	182	233	280
Paraguay	5	10	100	100	100	90	234	303	415	550
Peru	n.a.	n.a.	n.a.	100	100	.	.	.	2,205	2,710
Puerto Rico	10	3	7	6	6	235	81	224	211	234
Uruguay	5	100	100	100	.	128	279	296	316	.
Venezuela	2	22	100	30	100	132	1,060	1,452	1,802	2,420
Total						9,784	17,353	20,824	39,436	48,314

. = No census taken in this decade; no sample possible.

n.a. = Microdata incomplete or missing, but census was taken.