IPUMS-International Integrated Census Microdata Extract System: Users and Uses, May 2002-March 2007

Robert McCaa, Steven Ruggles, and Matt Sobek
University of Minnesota Population Center
The IPUMS-International projects receive major funding from the National Science Foundation of the United States, grants sbr-9908380 and ses-0433654.

"I have recently taken on a new job at [an international organization] and will be supporting the developing countries in carrying out their censuses. I am therefore interested to understand more about what data have been collected in the past across countries."
--project description, application #1759, www.ipums.org/international

Abstract. Launched in Beijing seven years ago at the 19th ANCSDAAP conference, the IPUMS-International census microdata integration project has flourished thanks, in no small part, to the support of National Statistical Offices (NSOs) of Asia and the Pacific region. This paper reports on accomplishments to date with respect to microdata anonymization, integration and dissemination, with particular emphasis on users and uses of the IPUMS database.

Briefly, sixty-seven NSOs have endorsed the project Memorandum of Understanding with the University of Minnesota. Fifty-eight have entrusted microdata to the University for a total of 172 censuses. Instead of disseminating source files entrusted by NSOs, the IPUMS project integrates census samples using composite coding, variable-by-variable. Currently integrated are samples for 63 censuses, 20 countries, and 185 million person records. Asia and the Pacific region account for about one-fourth of the database. Sixteen countries have endorsed the project MOU, 15 have entrusted data, and 4 are integrated—Cambodia, China, Philippines, and Vietnam. While data for thirty-nine censuses for the region have been entrusted to the University, only seven have been integrated to date. The small number is due in part to the difficulty of acquiring complete sets of data and documentation. Fortunately these problems are now resolved for Indonesia and Malaysia, and hopefully will soon be resolved for Fiji Islands, Mongolia, Pakistan, Thailand, and a number of other countries in the region.

In 2006, the IPUMS integrated metadata system was launched. By means of 5 "clicks" it is now possible to compare the phrasing of any census question for any combination of countries and censuses in the database. Only 3 clicks are required to view original source documentation in the official language or in English translation. With respect to users, of 1,763 completed applications for access, 1,264 undertakings (72%) were approved to gain access to the database. The report discusses usage statistics in detail.

Looking ahead, the biggest challenge for the IPUMS collaboratory will be the integration of 2010 round census microdata samples: to zealously protect statistical confidentiality, attain the highest standards of integration, and manage access to extracts of samples by researchers around the globe in a timely way and at no cost—all accomplished with a minimum of delay.

1. What are census microdata? Census *microdata* provide information about individual persons, 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. Microdata are exceedingly useful because they allow researchers to interrelate any desired set of population and housing characteristics (Dale, Fieldhouse and Holsworth, 2000). The flexibility offered by microdata is essential for comparative research because aggregate tabulations are often 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).

2. Integrated, high-precision samples. The IPUMS-International project is a global collaboratory of universities, National Statistical Offices, and international research institutes to preserve, integrate and manage access to 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 National Statistical Offices (NSOs) in sixty-seven countries, encompassing more than sixty percent of the world's population (Table 1). Fifty-eight NSOs have entrusted microdata to the project for a total of 172 censuses.

Table 1 near here

The project does not disseminate census files entrusted by NSOs. Instead high-precision census samples are anonymized (McCaa et. al. 2006) and integrated, variable-by-variable, using a composite coding system (Esteve and Sobek, 2003). In May 2002, the first phase of the project concluded with the integration of four census samples for Colombia, five for France, two for Kenya, four for Mexico, five for the United States, and two for Vietnam. In 2003, one sample for China (1982) was launched and in 2003, five samples for Brazil (1960-2000). In 2006, data for twelve additional countries (35 censuses) were integrated into the database: Belarus (1), Cambodia (1), Chile (5), Costa Rica (4), Ecuador (5), Greece (4), Philippines (3), Romania (2), South Africa (2), Spain (3), Uganda (2), and Venezuela (3). Currently, integrated samples may be accessed for 20 countries, 63 censuses and 185 million person records. Densities for most of the samples are ten percent, although some are five percent and a few are even less.

Asia and the Pacific region account for about one-fourth of the database. Sixteen countries have endorsed the project MOU, 15 have entrusted data, and 4 are integrated—Cambodia, China, Philippines, and Vietnam. While data for thirty-nine censuses for the region have been entrusted to the University, only seven have been integrated to date. The small number is due in part to the difficulty of acquiring complete sets of data and documentation. Fortunately these problems are now resolved for Indonesia and Malaysia, and hopefully will soon be resolved for Fiji Islands, Mongolia, Pakistan, Thailand, and a number of other countries in the region. Negotiations are underway with the census authorities of Bangladesh, India, Nepal and a number of other NSOs in the region.

Over the next three years, thanks to sustained funding by the National Science Foundation and the National Institutes of Health, samples for more than 80 censuses (25-

30 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.

3. Five (5) Clicks to compare any question in any combination of countries and censuses: the IPUMS Dynamic Metadata (Documentation) System. To use census microdata well, researchers must consult original source documentation to understand census concepts, definitions and nomenclatures. IPUMS has developed a dynamic metadata system to facilitate comparison of the phrasing, in English, of any census question with any combination of countries and years in the database. Five clicks are all that is required. For example, if the researcher wishes to compare the marital status question in the censuses of Mexico with those of the United States and Spain, this is accomplished simply by (1) accessing the web page (www.ipums.org/international), selecting (clicking) (2) "Variables", (3) the countries and census samples (years), (4) the marital status variable, and finally (5) "enumeration text". At this point, a screen will appear, displaying the census question, categories, and enumerator instructions—all in English. If the researcher wishes to see the sample frequencies of the codes for any variable, from that variable page (click "4", above), click codes, then "Case-count view" and the frequencies will appear for the selected variable and census samples. For complex variables, click "detailed codes" to see the full range of codes in the IPUMS database for this variable. Images of the original forms and instructions in the official language(s) of each census are also available, by means of 2 clicks from the home page (click "Census Questionnaires", then the specific country, census year and document). Images are not "book-marked", therefore, unlike for the dynamically generated enumeration text pages, scrolling is required to navigate to the item of interest.

Using the dynamic enumeration text pages, the researcher may easily study the displayed documentation to determine if the phrasing of questions in the various censuses is sufficiently alike to permit comparison, given the precise topic of research. For example, in Spanish censuses as recently as 2001, marital union was limited to legal status. In contrast, as long ago as 1960, Mexican censuses included consensual union as an option on the census form. If the researcher is studying consensual unions, the censuses of Spain, which define marital status as legal civil status, are not adequate to directly address this subject. Consider however that the 2001 microdata sample of Spain in the IPUMS database contains an imputed consensual union status variable ("Type of couple"), which may be suitable for some research questions.

Note that no registration is required to use the IPUMS dynamic enumeration text feature. Unlike access to the microdata, any user may browse IPUMS census documentation without registration. However, dynamic enumeration text is available only for censuses for which the microdata are integrated into the IPUMS database. It is for this reason that when microdata are entrusted to IPUMS, a concerted effort is made to obtain complete documentation—forms, instructions, codebooks, and technical or methodological studies—so that these may be integrated into the metadatabase. For non-English language materials, translators of all the major world languages are contracted by the IPUMS project to provide English translations of source documentation.

For the 2010 census round, some of our NSO partners are considering providing "tagged" census metadata so that they can be transferred to the IPUMS system with minimal additional treatment. We would be delighted to discuss this exciting development in detail with delegates of interested NSOs.

4. Managing Access ("Extracts"). Researchers must first be approved before access to any microdata is permitted. Moreover users are never permitted access to the original source files provided by the NSOs. Instead, data are provided 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.

In 2006, the Economic Commission of Europe published guidelines for Managing Statistical Confidentiality and Microdata Access. An IPUMS-Europe case study, using the specific case of France, is appended to the UNECE report as an example of good practice (Table 2). The case study describes how IPUMS manages access to microdata, explains why it is a good practice, identifies the target audience, explains confidentiality measures, specifies the rules and procedures regarding user access, summarizes supporting legislation (for others see http://unstats.un.org/unsd/goodprac/default.asp), and lists strengths and weaknesses as well as bibliographical references. While the IPUMS case study is European in scope, the details are nearly identical for the International project.

Table 2 near here

To request an extract, the researcher must first sign in by entering the registered password. To create an extract, the user makes a series of selections—country (or countries), census years, samples, variables and sub-populations—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.

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 use them with any statistical software.

5. Applicants and Approved Users. The IPUMS-International project offers bona fide researchers custom-tailored extracts at no charge via the Internet. During the first 58 months of operation (May 2002-March 2007), 1,763 applications were received, of which 72% were approved. The principal reason for rejecting an application is that the proposed research 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. 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. Approval is based solely on criteria of scientific feasibility (that census microdata are essential for the proposed research), including the credentials of the researcher.

The following statistics are derived from 1,264 undertakings approved for access to the IPUMS-International database. (Note that this is separate from the IPUMS-USA database, which is open-access and has over 23,000 registered users.) Incomplete applications are not included in this count. Nor is any supplemental information considered which may have been requested from applicants in weighing a decision on whether to grant access or not.

6. Who uses the data? The succinct answer to the question of who uses the IPUMS database is university professors and students. At 91%, they account for almost the entire group of users. Nevertheless, access has been granted to 31 researchers affiliated with international agencies, such as the World Bank, International Monetary Fund, DFID, etc. Twenty-six users are affiliated with international research institutes and 21 with United Nations organizations (ILO, WHO, UNSD). Eighteen are official statisticians. An equal number are national government employees. To date only three researchers affiliated with Non-Governmental Organizations have been granted access to the database.

Hundreds of university and research institutes are represented among IPUMS users. China is a good example. Fifteen centers are listed as the primary affiliation by Chinese users, as follows: China Center for Economic Research, Peking University; Chinese Academy of Social Sciences; Guanghua School of Management, Peking University; Institute of Policy and Management, Chinese Academy of Sciences; Institute of Population and Labor Economics, Chinese Academy of Social Sciences; Jinan University; Management and Administration Institute; Nankai University; National Bureau of Statistics of China; Renmin University of China; Sociology Department, Nanjing University China; The Institute of Population Research, Peking University; Tsinghua University; University of International Business and Economics; University of Science and Technology of China

Economists account for 44% of users, followed by demographers (13%), sociologists (12%), public policy analysts (5%), and historians (4%). A miscellany of 32 disciplines combined accounts for the remaining 22% of users.

Fifty-five countries are represented among the users. Four of five reside in countries whose samples are already entrusted to the University of Minnesota. Of all users, the

United States accounts for the largest percentage (64%), followed by the United Kingdom (4%), Colombia, Brazil and Canada (3% each), and Mexico, France, Spain and Germany (2% each). Ten countries account for 10% of users (roughly 1% each): Australia, Austria, Belgium, China, Italy, Japan, Kenya, Netherlands, Singapore, and Switzerland. One-fifth of researchers reside in countries not represented in the database—12% in countries where the data are entrusted to IPUMS but not yet integrated, and the remainder in countries that have not yet endorsed the IPUMS Memorandum of Understanding. It is remarkable that 8% of researchers use the database even though the availability of samples for their country has not progressed beyond an introductory phase—as is the case for Australia, Hong Kong, India, Japan, New Zealand, Republic of Korea, Russia, and Singapore.

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, UK and Canadian universities are nationals using the IPUMS-International database to study their country of origin, including not only countries in the developing world, but also the developed, such as France and Spain. .

7. Countries of research interest. The single most requested country, of those not integrated into the IPUMS database, is India (58 users), followed by Japan with 39. Ten or more users requested data for each of seven countries: Australia, Bangladesh, Indonesia, Republic of Korea, Pakistan, Russia, and Thailand. Of these, integration of samples is likely to be completed within a year or two for Indonesia, Pakistan and Thailand. In addition data were requested for Hong Kong, Iran, Laos, Macao, Myanmar, New Zealand, Singapore, and Sri Lanka. I would welcome the opportunity to discuss participation with delegates from countries whose census microdata have been requested, but not yet entrusted to the project.

8. Research topics. The epigraph indicates a somewhat surprising use of the database—a funding agency proposes to examine the content of censuses in preparation for the 2010 round of questionnaires. Most researchers use the data to address substantive economic, demographic and social issues. Given the prevalence of economists as users, the predominance of economic topics and econometric analysis should not be surprising, including such classics as the comparative study of labor force participation, demand and supply of public services (water, electricity, sewage, etc.), economic impact of family planning and fertility decline, discrimination in credit markets, econometric analysis of labor force and income, effect of long-term youth unemployment, effects of volume of human capital on returns to education, human capital and aging, impact of trade policies on growth, development, immigration, labor markets, and inequality, development of a system of regional accounts, transformation of labor market structure encompassing child and informal labor, etc.

Demographic topics include fertility analysis using the own-child method, infant and child mortality, population forecasts taking into account education, marital status, age, and sex, projection of housing needs, projection of educational needs, comparative study of immigrant groups in developed countries, effects of emigration on labor markets, effects of immigration on wages, home ownership and immigration, effects of immigration on poverty and economic welfare, scope and scale of international retirement migration, effects of urbanization on settlement patterns, etc.

Sociologists are studying comparative gender segregation in employment, gender diversity in agricultural economies, gender gaps in education and marriage, poverty and social issues, age hypogamy in marriage, social correlates of poverty, multidimensional measurements of poverty, etc. Public health specialists are researching the world health workforce, and the global impact of disease. 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.

Official statisticians are interested in census taking, international comparison of census data, and using census data to develop statistical methods. One researcher is studying "IPUMS methods for metadata and microdata dissemination and managing access to large amounts of data and documentation". 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 in Asia and the Pacific Region, integration of census microdata is an important 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 67 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 are likely to increase by an order of magnitude to become the most widely used demographic database in the world.

References.

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

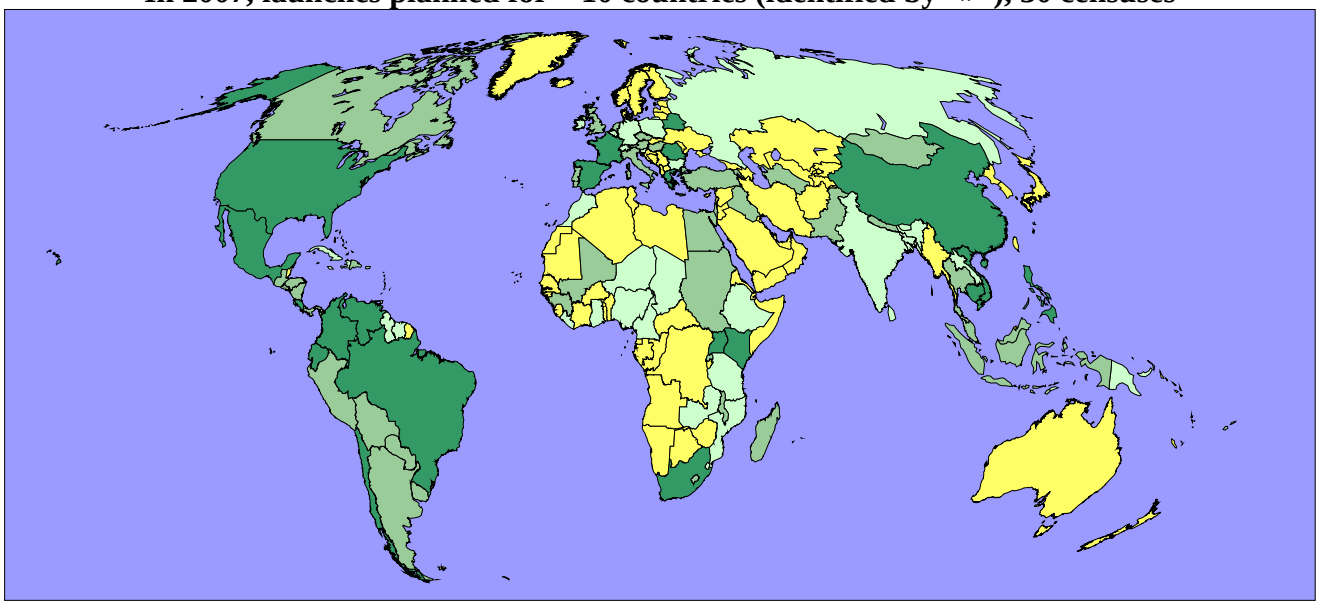
Esteve, Albert and Matthew Sobek. (2003). Challenges and Methods of Census Harmonization. *Historical Methods* 36: 66-79.

McCaa, 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.

McCaa, Robert, Steven Ruggles, Michael Davern, Tami Swenson, Krishna Mohan Palipudi. 2006. "IPUMS-International High Precision Population Census Microdata Samples: Balancing the Privacy-Quality Tradeoff by Means of Restricted Access Extracts," Privacy in Statistical Databases. Berlin: Springer, pp. 375-382.

Ruggles, Steven, Miriam King, Deborah Levison, Robert McCaa, and Matthew Sobek. (2003). "IPUMS-International: An Overview". *Historical Methods*, 36: 60-65.

Table 1. IPUMS-International Inventory: census microdatasets entrusted to project by country
Note: in 2006 there were 2 launches of integrated microdata totaling 11 countries and 35 censuses
In 2007, launches planned for ~10 countries (identified by "»"), 30 censuses



Key: *dark green*: disseminating; *medium green*: data received; *lightest green*: negotiating; *yellow*: no desire to participate
bold country = Memorandum of Understanding signed with Regents of the University of Minnesota
Year = census conducted; ***Bold year*** = microdata survive; * = 100% micordata entrusted, where extant; *m* = microcensus

No. of samples and densities			Country	2000s	1990s	1980s	1970s	1960s
10%	5%	<=4%						
Release 1, 2002/3 (28 samples)								
5			**Brazil ('60 recovered)**	**2000**	**1991**	**1980**	**1970**	**1960**
		2	**China ('90 in process)**	**2000**	**1990**	**1982**		1964
3		**1**	***Colombia ('05 in preparation)**	**2005**	**1993**	**1985**	**1973**	**1964**
	5		**France ('99 in preparation)**	**1999**	**1990**	**1982**	**1975**	**1968, 62**
	2		**Kenya ('79 & '69 in preparation)**	**1999**	**1989**	**1979**	**1969**	
5		**2**	**Mexico ('80 2/3 recovered)**	**2000, 05**	**1990, 95**	**1980**	**1970**	**1960**
	5		**United States ('05 in preparation)**	**2000, 05**	**1990**	**1980**	**1970**	**1960**
	2		**Vietnam ('89 recovered)**		**1999**	**1989**	1979	
Release 2, May 2006 (19 samples)								
4		**1**	***Chile ('60-82 recovered)**	**2002**	**1992**	**1982**	**1970**	**1960**
3	**1**		***Costa Rica ('63-84 recovered)**	**2000**		**1984**	**1973**	**1963**
4		**1**	***Ecuador ('60-'82 recovered)**	**2001**	**1990**	**1982**	**1974**	**1962**
2			**South Africa ('70-'91 omitted)**	**2001**	**1996, 91**	**1985, 80**	**1970**	1960
4			***Venezuela ('01, '61 in process)**	**2001**	**1990**	**1981**	**1971**	**1961**
Release 3, December 2006 (16 samples)								
1			**Belarus ('89 lost)**		**1999**	1989	1979	1970
1			**Cambodia**		**1998**			1962
4			**Greece ('71 recovered)**	**2001**	**1991**	**1981**	**1971**	1961
3		**2**	**Philippines ('60, '70 recovered)**	**2000**	**1990**	**1980**	**1970**	**1960**
3			**Romania ('77 recovered)**	**2001**	**1992**		**1977**	1965
	3		**Spain ('81 recovered)**	**2001**	**1991**	**1981**	1970	1960
2			***Uganda ('80 is incomplete)**	**2002**	**1991§**	1980§		1969

Europe (33 datasets, including datasets in dissemination listed above)								
4			**Austria ('61 lost)**	**2001**	**1991**	**1981**	**1971**	1961
			Bulgaria (in process)	**2001**	**1992**	**1985**	1975	1965
	2		**Czech Republic ('70 recovered)**	**2001**	**1991**	**1980**	**1970**	1961
			Germany (in process)	**2001m**	**1991m**	**1987, 81**	**1970, 71**	1961
	4		**»Hungary ('70 recovered)**	**2001**	**1990**	**1980**	**1970**	
			Italy (in process)	**2001**	**1991**	1981	1971	1961
		3	**Netherlands ('60 recovered)**	**2001m**			**1971**	**1960**
	3		**»Portugal ('70 not recoverable)**	**2001**	**1991**	**1981**	1970	1960
			Slovenia (recovery underway)	**2001**	**1991**	**1981**		
			Switzerland (4-5% in preparation)	**2000**	**1990**	**1980**	**1970**	1960
			Turkey (recently signed)	**2000**	**1990**	**1980**, 85	1970, 75	1960, 65
		2	**»United Kingdom (in process)**	**2001**	**1991**	**1981**	**1971**	**1966, 61**
North America and the Caribbean (39 datasets; includes datasets in dissemination listed above)								
		4	**»Canada**	**2001**	**1991,** 96	**1981,** 86	**1971,** 76	1961, 66
			Dominican Republic (in process)	**2003**	**1993**	**1981**	**1970**	1960
1			***El Salvador ('71 recovered)**		**1992**		1971	1961
4		**1**	***Guatemala (64-81 recovered)**	**2003**	**1994**	**1981**	**1973**	**1964**
3		**1**	***Honduras ('61-88 recovered)**	**2000**		**1988**	**1974**	**1961**
2			***Nicaragua ('71 recovered)**	**2005**	**1995**		**1971**	1963
5			***Panama ('60-80 recovered)**	**2000**	**1990**	**1980**	**1970**	**1960**
	4		**Puerto Rico ('70 -80 recovered)**	**2000**	**1990**	**1980**	**1970**	1960
2			***Saint Lucia**	**2001**	**1991**	**1980**	1970	1960
South America (40 datasets; includes datasets in dissemination listed above)								
3		**1**	**»Argentina ('70 recovered)**	**2001**	**1991**	**1980**	**1970**	1960
3			***Bolivia ('76 recovered)**	**2001**	**1992**		**1976**	
4		**1**	***Paraguay ('62-82 recovered)**	**2002**	**1992**	**1982**	**1972**	**1962**
1			***Peru ('81 recovery uncertain)**	**2006**	**1993**	**1981**	1972	1961
4			***Uruguay ('63 recovered)**		**1996**	**1985**	**1975**	**1963**
Africa (21 datasets; includes datasets in dissemination listed above)								
2			**»*Egypt ('86 and 96 recovered)**	**2006**	**1996**	**1986**, 81	1976	1964
2			***Guinea, Conakry**		**1996**	**1983**		1960
			Lesotho (in progress)	**2006**	**1996**	**1986**	**1976**	1966
1			***Madagascar ('93 recovered)**		**1993**			
3			***Malawi ('77 recovered)**	2008	**1998**	**1987**	**1977**	1967
3			***Mali ('76 in progress)**		**1998**	**1987**	**1976**	
2			**Mauritius ('83 and '72 uncertain)**	**2000**	**1990**	**1983**	1972?	1962
2			**»*Rwanda ('91 recovered)**	**2002**	**1991**			
2			***Sudan ('73 recovery underway)**	2007	**1993**	**1983**	**1973**	
Asia and Oceania (39 datasets; includes datasets in dissemination listed above)								
1			**Armenia ('89 lost)**	**2001**		1989	1979	1970
			Australia (invited; '76 earlier lost)	**2001, 06**	**1991, 96**	**1981, 86**	1971, 76	1961, 66
			Bangladesh ('81 to be recovered)	**2001**	**1991**	**1981**	1974	1961
			Bhutan (invited)	**2005**				
3			***Fiji Islands ('76 in recovery)**	2007	**1996**	**1986**	1976	**1966**
			India (invited; '81 recoverable?)	**2001**	**1991**	1981	1971	
4	**3**		**Indonesia**	**2000, 05**	**1990, 95**	**1980, 85**	**1971**	1961
1			**»*Iraq ('87 destroyed by looting)**	2007	**1997**	1987	1977	1967

4			**»Israel (’72, 83 recovered)**		**1995**	**1983**	**1972**	**1961**, 67
			Japan (invited)	**2000, 05**	**1990, 95**	**1980, 85**	1970, 75	1960, 65
			Kazakhstan (invited)	**1999**	1989	1979	1970	1959
			Korea, DPR (invited)		**1993**			
			Korea, RO (invited)	**2000, 05**	**1990, 95**	**1980, 85**	**1970, 75**	**1960, 66**
			Laos, DPR (invited)	**2005**	**1995**	**1985**	1973	
		4	**»Malaysia (’70, ’80 recovered)**	**2000**	**1991**	**1980**	**1970**	1960
			Maldives (negotiating)	**2006, 00**	**1995, 90**	**1985**	1977, 74	1965
1			***Mongolia**	**2000**		1989	1979	1970
			Myanmar (invited)			1983		
			Nepal (invited)	**2001**	**1991**	1981	1974	1961
			New Zealand (invited, ‘76 earlier lost)	**2001, 06**	**1991, 96**	**1981, 86**	1971, 76	1961, 66
3			***Pakistan (’81 in recovery)**		**1998**	**1981**	**1973**	1961
1			**»Palestinian Authority**		**1997**			
			Sri Lanka (invited)	**2000**		**1981**	**1971**	
		4	**Thailand**	**2000**	**1990**	**1980**	**1970**	1960
			Timor Leste (invited)	**2004**				
1			**Turkmenistan (’89 lost)**		**1995**	1989	1979	1970

Table 2. Case Study for **Managing Statistical Confidentiality and Microdata Access: Principles and Guidelines of Good Practice**:
http://www.unece.org/stats/documents/tfcm/1.e.pdf

Case Study – Arrangements for Providing International and National Access to Anonymized Census Microdata Samples via the IPUMS-International and the Integrated European Census Microdata websites (University of Minnesota Population Center and the Centre d'Estudis Demogràphics, Autonomous University of Barcelona) with France, as a specific example.

1. Broad description

High precision, anonymized, integrated census microdata are available to researchers on a restricted access basis from IPUMS-International (www.ipums.org/international). Terms are specified by a memorandum of understanding negotiated between each National Statistical Office and the University of Minnesota. This method of dissemination is governed, on the one hand, by legislation requiring that the data be held in strict confidence and used exclusively for statistical purposes and, on the other, by a stringent license agreement between the University of Minnesota and each user. In May 2002, anonymized, integrated microdata samples for the French censuses of 1962, 1968, 1975, 1982 and 1990 were released, along with samples for China, Colombia, Kenya, Mexico, the USA and Vietnam. The December 2006 release includes samples for the censuses of Belarus, Greece, Romania and Spain as well as the Philippines and Uganda. As of January 1, 2007, the database comprises 63 samples, 20 countries, and 185 million person records. An additional six European statistical agencies (and 38 non-European) have provided census microdata to the project: Austria (4 censuses), Czech Republic (2), Hungary (4), Netherlands (3), Portugal (3), and the United Kingdom (2; the 1981 and earlier censuses are under consideration). Five other European countries have endorsed the project, but have not yet provided data: Bulgaria, Germany, Italy, Slovenia, and Turkey. Beginning in 2008, the European microdata will also be distributed by the Integrated European Census Microdata (IECM) project using identical protocols, although the microdata will be harmonized according to European, rather than global, practices.

2. Why is it a good practice?

Conditions of access are transparent and provide a degree of certainty to users and the National Statistical Offices. Sanctions for violations of misuse are clearly spelled out and enforceable by a set of strong administrative and legal mechanisms. The microdata are anonymized by means of a variety of technical measures, including the suppression of detailed geography. Variables are integrated using a composite coding scheme to facilitate temporal and cross-national comparative research. The documentation, including both scanned images of forms and instructions as well as integrated metadata, is extensive and available at no cost. The microdata are also available at no cost, but availability is restricted to approved academic and policy researchers. These practices are in compliance with the Fundamental Principles of Official Statistics.

3. Target audience

The research community, including academic and policy makers regardless of country of birth, residence, work-place or citizenship.

4. Detailed description

The IPUMS-International project is governed by a uniform Memorandum of Understanding (MOU) signed with each participating National Statistical Office. The MOU (copy appended below) confirms that the National Statistical Office specifies the

terms and conditions under which the microdata and metadata entrusted to the University of Minnesota and the Autonomous University of Barcelona shall be governed:

1) the NSO retains ownership, including copyright;
2) data are to be used exclusively for statistical purposes associated with teaching, research, and publishing;
3) use for administrative, commercial or income generating purposes is prohibited;
4) application procedures for obtaining access to microdata are specified in the MOU;
5) confidentiality of the data is protected by means of prohibitions against
    a. any attempt to ascertain the identity of individuals, families, households, dwellings or other identities
    b. any allegation that an identification has been made.

In addition there are statements regarding:

6) the necessity of security measures for retaining microdata;
7) publication and citation requirements;
8) procedure for dealing with violations, including sanctions;
9) the sharing of integrated microdata with the National Statistical Offices;
10) recognition of jurisdiction under international law with the ICC International Court of Arbitration for the settlement of disputes; and
11) establishing the supreme precedence of the MOU over any subsidiary document, contract or other instrument.

The principal sanction for misuse is recall of data and an embargo against use by the individual and the individual's institution. In addition, the sanctions clause of the MOU threatens additional sanctions to assure compliance:

"Violation of the user license may lead to professional censure, loss of employment, and/or civil prosecution. The University of Minnesota, national and international scientific organizations, and the [the Statistical Agency of Country X] will assist in the enforcement of provisions of this accord."

4.1 Data confidentiality

Before providing census microdata to the Minnesota Population Center, the National Statistical Office imposes a number of undisclosed technical confidentiality measures. The Minnesota Population Center imposes an additional suite of techniques such that any allegation that an individual has been identified with absolute certainty is false. In addition, to further ensure the confidentiality of the microdata, administrative geography is limited. In the case of France 22 regions are identified. The smallest has a population exceeding 80,000 in the 1990 census (sample n > 4,000). The sample count for any identifiable single year of age is >100. For any identifiable country of citizenship the sample count is >100. Each National Statistical Office determines the minimum population threshold for the identification of administrative geography and other sensitive characteristics, such as ethnicity, country of birth, citizenship, etc.

4.2 Rules and procedures regarding release to users

Prospective users must complete an electronic application to gain access to the data. The preamble of the application reads:

Legal notice: Submission of this application constitutes a legally binding agreement between the applicant, the applicant's institution, the University of Minnesota, and the relevant official statistical authorities. Submitting false, misleading or fraudulent information constitutes a violation of this agreement. Misusing the data by violating any of the conditions detailed below also constitutes a violation of this agreement and may lead to

professional censure, loss of employment, or civil prosecution under relevant national and international laws, and to sanctions against your institution, at the discretion of the University of Minnesota and the official statistical authorities."

The application form requires that the applicant indicate agreement, by electronically checking specifically each of eight conditions of use, including the following:

☑ **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. Statistical results that might reveal the identity of persons or entities may not be reported or published in any form.

And:

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

Violation of this agreement will lead to revocation of this license, recall of all microdata acquired, a motion of censure to the relevant professional organization(s) and civil prosecution under national or international statutes, at the discretion of the Regents of the University of Minnesota and the official statistical agencies. Sanctions likewise may be taken against the institution with which the violator is affiliated.

Failure to indicate agreement with any one of the conditions automatically disqualifies the applicant for access to the microdata. In addition the successful applicant must provide detailed information on academic qualifications, affiliation, research experience, source of funding, bona fides, and familiarity with human subjects protections regarding statistical confidentiality. Finally the applicant must submit a project description demonstrating need for access to census microdata. Applications are reviewed by senior principal investigators. Approximately 1/3 of applicants who complete the form are denied access. The application is valid for one year and may be renewed.

5. Supporting legislation (example of France)

Article 6 of the law of 1978 introduced the possibility for statisticians and researchers to use personal data, including nominative data, originally collected for purposes other than historical or scientific research or statistics. More precisely, it indicates that subsequent processing for statistical or research purposes is always compatible with the objectives for which the data had been collected. French Act no. 2004-801 of August 6, 2004 amends and updates the Statistics Law of 1978 to protect individuals with regard to the processing of personal data and the free movement of these data. The Act is in compliance with the European directive no. 95/46/CE of October 24, 1995 of the European Parliament and Council. Information on legislation regarding good practices is available at: http://unstats.un.org/unsd/goodprac/default.asp For information on statistical confidentiality, microdata access and privacy, see "Principle 6".

6. Strengths

a. Offers security against loss of source microdata. Raw data files entrusted to the project are encrypted and stored in a secure data repository. Copies of these files are made available only to the National Statistical Office-owner, and are never re-distributed to others.

b. Fosters maximum uniformity of approach and facilitates greater access to microdata by the research community.

c. Improves on arrangements for providing access to microdata to the greater satisfaction of both the National Statistical Offices and the research community.
d. National Statistical Offices cede census microdata files to the University of Minnesota. The data are anonymized and then integrated. Much new integrated metadata are written and stored in a database accessible to all at no cost via the internet. Integrated microdata are available for dissemination on a licensed basis to approved researchers. All licensed microdata disseminated by the University of Minnesota Population Center are governed by a uniform Memorandum of Understanding (MOU) between the National Statistical Office and the University. If requested to do so, the University will cease dissemination and return all copies of census microdata in its possession to the corresponding National Statistical Office.
e. Employees of the University who work with original source data are certified in human subject protections, including the protection of statistical confidentiality. Violations are punishable by termination of employment, and, at the discretion of the University, civil prosecution with a maximum fine of US$250,000 and/or three years imprisonment.
f. The means of gaining access to the microdata are transparent and equitable. They are based on the principle of freedom of scientific inquiry, regardless of country of birth, residence, workplace or citizenship. Decisions to grant access are determined by project principal investigators. Each individual who wishes to work with the microdata is required to be licensed. The license is valid for one year and is renewable. A condition for renewal is the sharing of research findings, which, in turn, are made available to the national statistical offices.
g. Microdata are available as extracts on a licensed basis only to researchers who agree to abide by the conditions of use and demonstrate a bona fide research need to access the data. The license constitutes a legally binding undertaking. An attempt to match individuals constitutes a violation of the license agreement and would lead to recall of data and sanctions against both the individual and his/her institution.
h. Sanctions for breaches of the license agreement are clearly spelled out. These include:
    i. sanctions against both the individual and the institution with which the individual is associated (e.g., University, international organization);
    ii. denial of access would immediately be invoked against the individual and his/her institution and would continue until corrective measures were deemed to be sufficient by the University of Minnesota and the National Statistical Office whose data were violated. If the institution where the breach occurred was the recipient of a grant from the National Institutes of Health of the United States, each researcher at the institution could be required to undergo Human Subjects Protection training and re-certification before access was re-instituted for individuals at that institution.
    iii. civil prosecution could be instituted with assistance requested, under the terms of the project MOU, of the National Statistical Office of the country in which the violation occurred to the extent permitted by national legislation.
i. Microdata are encrypted during transmission using 128-bit SSL (Secure Sockets Layer) encryption standards used by the financial industry.
j. Anonymization protocols (top coding, bottom coding, grouping of small cell counts, collapsing of variables, randomization of records and some recodes,

suppression of sensitive variables, etc.) are rigorous, yet precision of samples is high. Anonymization protocols are determined by each National Statistical Office before extracts of the data are disseminated.

k. Integrated metadata are provided describing census operations, sample methodologies, variables and codes. The documentation is harmonized so that researchers who become familiar with the metadata for one census will readily understand the metadata system for any other census of any other country.
l. Microdata consist of high precision household samples with many integrated, value-added variables—such as "WTPER", which specifies the person weight for each record in every sample; "SUBSAMP", which provides 100 certified sub-samples which researchers may use to generate robust estimates of sample variance; "SPLOC" which points to the spouse of each individual whose spouse in co-resident in a household; etc.
m. Costs are borne through sustained funding from the National Science Foundation of the United States of America with supplementary funds provided by the National Institutes of Health. Where required, the project pays a license fee to the National Statistical Office for the documentation and microdata. The fee is intended to cover marginal costs for the National Statistical Office to provide technical assistance in developing the microdata samples and interpreting the documentation. The **European Union Sixth Framework Programme** provides support to the IECM project for enhancing, harmonizing and disseminating the integrated European microdata and metadata as well as for coordinating tasks based in Europe.

7. Weaknesses

a. National Statistical Offices cede authority to the University to grant access to census microdata extracts to bona fide researchers. Decisions to grant access are determined by project principal investigators.
b. Microdata are not wholly anonymized. With sufficient resources, in terms of computing power, time, and a companion microdataset, data matching could be performed to identify individuals to a high probability, although not with absolute certainty.
c. Misuse of microdata by even one researcher may impact negatively on the ability of a National Statistical Office to obtain cooperation of respondents in that country, or even conceivably, other countries.
d. Users do not have access to original source files supplied by the National Statistical Office. Instead researchers access integrated microdata with codes and documentation which not only may differ from the original source but also may contain errors introduced in the integration process.
e. Quality of microdata may not be sufficiently high for the intended research purpose.
f. Whether the license constitutes a legally binding undertaking has not been tested in a court of law.
g. There is no requirement that the microdata be destroyed once the initial research is completed.
h. There is no opportunity for the National Statistical Office to comment upon the research before it is published.

8. References

Bruengger, Heinrich. 2004. "The relationship between the fundamental principle on confidentiality and population censuses: Statement from the UNECE Statistical

Division," United Nations Symposium on Population and Housing Censuses: New York, September 13-14.

Isnard, Michel. 2006. "Statistics and individual liberties: recent changes in French law," Courrier des statistiques, English series no.12, pp. 26-30.

McCaa, Robert and Steven Ruggles. 2002. "The Census in global perspective and the coming microdata revolution," Scandinavian Population Studies, 13:7-30.

McCaa, Robert and Wendy L. Thomas. 2003. "Archiving Census Documentation and Microdata: Preserving Memory, Increasing Stakeholders", Notas de Población XXIX(75):303-320

McCaa, Robert and Albert Esteve. 2006. "IPUMS-Europe: Confidentiality measures for licensing and disseminating restricted access census microdata extracts to academic users," Monographs of official statistics: Work session on statistical data confidentiality. Luxembourg: Office for Official Publications of the European Communities, pp. 37-46.

McCaa, Robert, Steven Ruggles, Michael Davern, Tami Swenson, Krishna Mohan Palipudi. 2006. "IPUMS-International High Precision Population Census Microdata Samples: Balancing the Privacy-Quality Tradeoff by Means of Restricted Access Extracts," Privacy in Statistical Databases. Berlin: Springer, pp. 375-382.

McCaa, Robert, Steven Ruggles, Matt Sobek, and Albert Esteve. 2006. Using integrated census microdata for evidence-based policy making: the IPUMS-International global initiative, African Statistical Journal, 2(May):83-100.

Letter of Understanding
Integrated Public Use Microdata Series International
and the **[National Statistics Institute of Country X]**

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

1. Ownership. The **[National Statistics Institute of Country X]** is the owner and licensee of the intellectual property rights (including copyright) in the metadata and microdata of [Country X] acquired by the University of Minnesota to be distributed by **Integrated Public Use Microdata Series International**. This agreement explicitly authorizes release to the University of microdata of [Country X] that may be in the possession of third parties. The University is obligated to provide to the **[National Statistics Institute of Country X]** timely notice of any such acquisitions and, upon request and without cost, provide copies of same.
2. Use. These data are for the exclusive purposes of teaching, scientific research and publishing, and may not be used for any other purposes without the explicit written approval, in advance, of the **[National Statistics Institute of Country X]**.
3. Authorization. To access or obtain copies of integrated microdata of [Country 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 Statistics Institute of Country X]**, or other authorized distributors. Once approved, the user is licensed to acquire integrated

metadata and microdata of [Country 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 [Country X] is permitted in communications such as scholarly papers, journals and the like. The authors of these communications are required to cite **[National Statistics Institute of Country X] and Integrated Public Use Microdata Series International** as the sources of the data of [Country X], and to indicate that the results and views expressed are those of the author/user.
8. Violations. Violation of the user license may lead to professional censure, loss of employment, and/or civil prosecution. The University of Minnesota, national and international scientific organizations, and the [National Statistics Institute of Country X] will assist in the enforcement of provisions of this accord.
9. Sharing. **Integrated Public Use Microdata Series International** will provide electronic copies to the **[National Statistics Institute of Country X]** of documentation and data related to its integrated microdata as well as timely reports of authorized users.
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 arbitrator, who shall be selected by the ICC International Court of Arbitration. This agreement shall be governed by, and construed in accordance with, generally accepted principles of International Law.
11. Order of Precedence. In the event of a conflict between a term or condition of this Letter of Understanding and a term or condition of any Contract, to which this Letter of Understanding is attached, the term or condition in this Letter of Understanding shall prevail.

Date: _____

Signed: _____
Regents of the University of Minnesota
By: Kevin J. McKoskey, Sponsored Projects Administration

Date: _____

Signed: _____
Rev. Jan. 27, 2005