IPUMS-International: Free, Worldwide Microdata Access

Now for Censuses of 62 Countries--80 by 2015

58th International Statistical Institute, Dublin, Ireland, 21-26 August, 2011

McCaa, Robert
University of Minnesota Population Center
50 Willey Hall, 225 19th Ave S.
Minneapolis, MN 55455 USA
E-mail: rmccaa@umn.edu

Ruggles, Steven
University of Minnesota Population Center
50 Willey Hall, 225 19th Ave S.
Minneapolis, MN 55455 USA
E-mail: ruggles@umn.edu

Sobek, Matthew L.
University of Minnesota Population Center
50 Willey Hall, 225 19th Ave S.
Minneapolis, MN 55455 USA
E-mail: sobek@umn.edu

Thomas, Wendy
University of Minnesota Population Center
50 Willey Hall, 225 19th Ave S.
Minneapolis, MN 55455 USA
E-mail: wlt@pop.umn.edu

"Dissemination [means] opening up the value inherent in our data"

Seminar on Emerging Trends in Data Communication and Statistics, New York Feb. 19, 2010

Walter Radermacher (President, Eurostat) and Pieter Everaers (Director, Eurostat)

ABSTRACT.

The Minnesota Population Center (MPC), through the IPUMS-International census microdata project, archives the world's largest stock of census microdata and documentation. A decade of labor assiduously scouring local, national, regional, and international archives on every continent is beginning to bear fruit. Microdata for over 350 censuses for more than 120 countries are safely ensconced in the MPC digital archives. Metadata from more than 900 censuses have been catalogued, scanned and are now being disseminated world-wide without cost in cooperation with National Statistical Institute (NSI) partners and the Integrated Health Survey Network, using the latest international standards for electronic metadata. 5,000 researchers representing more than 80 countries are registered to access confidentialized, integrated microdata without payment and with complete

academic freedom—thanks to a uniform licensing agreement endorsed by more the 90 NSIs. Integration lowers the barriers to entry and facilitates comparative research over space and time.

For the future, we plan to integrate and disseminate confidentialized samples of the 2010 round censuses of the 60 countries already represented in the database. Samples for an additional 20-30 countries will be released to the global scientific community as time and resources permit. New initiatives are also planned: boundary files for GIS applications, an on-line tabulator for registered researchers, a secure enclave offering access to full-count microdata at the MPC and perhaps virtual enclaves for partners with certified secure sites. Several NSI partners have already granted assent for constructing a pilot at the MPC. Before the end of this year, thanks to major funding from the National Science Foundation (USA), a new project, TerraPop, begins--an initiative to combine population microdata with climate and land cover data.

Keywords. *census microdata, microdata access, integration, dissemination,*

PAPER.

For population census microdata access, the future is now at IPUMS-International, www.ipums.org/international. From June 2011, 185 high-precision, confidentialized, integrated samples representing 62 countries and totaling 397,316,462 person records are available to researchers free of charge (Table 1). The number of users and usage is commensurate, as we will illustrate below, with the scale of the database (and the scale of two decades-long, sustained investment in social science microdata infrastructure by the National Science Foundation and National Institutes of Health of the USA).   The microdata encompass 70% of the world's population. Each year samples for an additional five to seven countries are integrated into the database.   For 2011, these include Germany (4 censuses), Ireland (8), Jamaica and Malawi (3 each), Iran, Sierra Leone, and Sudan (1 each).

Samples for the 2010 round of censuses are assigned the highest priority to make them available to researchers from the IPUMS-I website within two or three years of enumeration day. In this regard, we are especially grateful to the General Statistics Office of Vietnam for entrusting a mere 18 months after enumeration day the long-form microdata for the population census 2009, the Central Bureau of Statistics of Sudan (2008 long form census data for both North and South), the National Institute of Statistics and Economic Studies of France (2004-8), the Statistical Centre of Iran (2006), the National Statistics Institute of Cambodia (2008), and the National Statistics Office of Malawi (2008).   Thanks to this cooperation in facilitating copies of metadata and microdata, it was possible to fast-track integration into the IPUMS-I database for official launch at the 58th ISI meeting in Dublin. June 2012, the 2010 census samples of Indonesia and Mexico are scheduled for launch—precisely because the data as well as comprehensive documentation were made available within a mere year of census day.   National Statistical Office partners of the IPUMS-International project are encouraged to entrust copies of 2010 round microdata and metadata in a timely fashion to avoid delay in the integration and launch process.   What is required for efficient, speedy integration is explained in our paper presented at the UNECE census expert group meeting "Census Outputs to Meet User Needs" in Geneva two years ago (McCaa and Esteve 2009).

In the spirit of the epigraph—the President of Eurostat's injunction to open "up the value inherent in our data"—by June 2015, the IPUMS-I database will host high precision household samples for approximately 85% of the world's population (80 countries), once the 292 census datasets already entrusted are processed.   Thanks to the cooperation of official statistical offices of 96

countries (Figure 1), a uniform memorandum of understanding specifying common agreement to eleven principles—ownership, use, access, restrictions, confidentiality, security, publication, violations, sharing, jurisdiction and precedence—governs access to the microdata (Conference of European Statisticians, 2007).   The 10 most populous countries yet to embrace the IPUMS-International principles are the Russian Federation, Japan, Congo (DR), Myanmar, Korea (RO), Algeria, Afghanistan, Uzbekistan, Saudi Arabia, and Korea (DR). Statistical offices not currently cooperating in the IPUMS-I initiative are cordially invited to consider doing so by contacting the first author of this paper.

Our paper briefly describes the IPUMS-International road map that got us to where we are and points to where we are going.   The paper is divided into five short sections: archiving, access, usage, integration of microdata and metadata, and future initiatives.

Archiving.   There is no future for microdata without the past. The Minnesota Population Center (MPC), through the IPUMS-I census microdata project, archives the world's largest stock of census microdata and documentation.   A decade of labor assiduously scouring local, national, regional and international archives around the globe is beginning to bear fruit (McCaa and Thomas 2009).   Microdata for over 350 censuses for more than 120 countries are safely ensconced in the MPC digital archives.   Metadata from more than 900 censuses have been catalogued, scanned, and are now being disseminated world-wide without cost.

In early 2011 IPUMS-I completed a project with the International Household Survey Network (IHSN) with funding by PARIS21 to generate metadata—country-by-country—for both integrated samples and for the original files as entered into the IPUMS-I database for harmonization. The metadata was structured to be used with the IHSN Microdata Toolkit, developed by the World Bank, which has been introduced in over 80 developing countries to promote the adoption of international standards and best practices for microdata management. The Toolkit documents data in accordance with the international standards of the Data Documentation Initiative (DDI) and Dublin Core. The metadata files created in this project were repatriated to the countries of origin along with PDF copies of major technical documents. In addition, copies were entered in the National Data Archive (NADA) catalog to provide broader access to the fully searchable content of the metadata files and direct researchers to IPUMS-International resources.

As part of this project, IPUMS-I has mapped its metadatabase and related metadata collection to the DDI standard structure. This allows us to generate DDI output customized to individual data extractions. The DDI can then be rendered as a PDF codebook or be used as input to a growing number of analysis tools that are able to understand and use the DDI structure.

In addition, the MPC can leverage a number of metadata creation and management tools to supplement its own in-house software development. It increases our flexibility and interoperability with systems outside of the MPC such as the NADA catalog and the DataVerse Network, an open-source application for publishing, citing and discovering research data.

Access. Access to the IPUMS-International microdata is restricted—despite the "P" in IPUMS. Would-be users must submit a detailed electronic application both to establish research bona-fides and to explain need for access.   An essential part of the application is to agree to ten stringent restrictions on condition of use—prohibiting redistribution, restricting to scholarly use, prohibiting commercial user, protecting confidentiality, assuring security, enforcing strict rules of confidentiality, permitting scholarly publication, citing properly, threatening disciplinary action for violations, and the reporting of errors.   In other words, the IPUMS-I is a "trusted user" access system.

The application binds both the researcher and the researcher's institution.   The Legal Counsel

of the University of Minnesota is posed to strike at the first indication of misuse to sanction not only the researcher, but also the researcher's institution.   Despite these restrictions almost five thousand researchers—representing 94 countries and over 800 institutions—are approved for access to the IPUMS-I database.   More than one-third of IPUMS-I trusted users request access to microdata for a single country. A large fraction of these are resident abroad and seek access to data for their own country of identity.

A mirror site for Integrated European Census Microdata (IECM) was inaugurated in 2008 at the Center for Demographic Studies (Autonomous University of Barcelona) and a second site for Africa (AICMD) in 2010 at the African Centre for Statistics.   Both sites emphasize their comparative advantage by disseminating specialized metadata and microdata for their respective regions.   The IECM site offers a European-flavored harmonization, an optimized version of IPUMS-International, that takes into account census principles and practices in the European region.   In addition the IECM project offers the first fully functioning cross-national tabulator of integrated census microdata. The ACS site offers access to African microdata, and, in addition, hosts on one convenient page an entire collection of original source census documents, county-by-country and census-by-census.

Usage. Usage of the IPUMS-I database in terms of sheer scale is astonishing.   24,699 extracts totaling 85,505 samples and 891,267 variables have been made to date.   From June 2010 to April 2011, the rate of increase in number of users is 25%; extracts, 45%; and variables extracted, 52%. Note, however, that the mean extract consists of microdata for a mere 1.8 countries, 3.5 samples, and 10.4 integrated variables.   The typical (median) user makes three extracts, consisting of four samples for one country and 19 variables. The top 5% of users, request 36 or more extracts, 26+ countries, 52+ samples and 110+ integrated variables.   The wonder of the web is that both these "power users" and novices may be serviced equally well by a dynamic metadata system and microdata extract engine at no significant additional cost.

These statistics may strike an odd-note to the ear of the official statistician accustomed to thinking in terms of static samples, where an identical, complete set of variables and metadata is disseminated to each user, regardless of need or level of experience.   The future of microdata is with web 2.0--dynamic metadata and dynamic extracts, where no two experiences are alike.   All the microdata products disseminated by the Minnesota Population Center (MPC), including IPUMS-I, are dynamic.   To obtain IPUMS-I microdata, the researcher must first log-in by means of a password to place an electronic order ("create an extract").   The next step is to select samples and variables by browsing the corresponding webpages.   To review selections, click the data cart.   Once the selections are complete, then proceed to make the extract ("check-out").   During the check-out process, a number of options are presented to refine the extract, including attaching characteristics, customizing sample size, etc.   Once the order is submitted, the extract engine generates a custom-tailored set of microdata and the corresponding metadata.   The user then logs-in, downloads the extract consisting of both metadata and microdata. and analyzes the extract with whatever hardware and software the researcher may wish to use.

Researchers report publication on the MPC "Bibliography" page.   The page is publicly available and includes citations of articles, books, dissertations, conference proceedings, and policy papers.   Click "IPUMS-International" to restrict the search to publications using IPUMS-I samples.

As noted above, the usage statistics reveal a surprisingly low average number of variables per extracts.   This is because most researchers are parsimonious, requesting only a few variables of specific interest for a research problem.   Likewise, the number of samples and countries per extract is also low because most researchers are interested in only one or two countries and three or four samples. Nonetheless there is a core of dedicated power users, who make a dozen or more extracts per

year on a wide range of samples, countries and variables. The list of the top forty institutions in terms of extracts includes many of the world's premier universities and research organizations (see Table 2).

The IPUMS-I "Top 40" institutions in terms of data usage are scattered across fourteen countries. In 46 countries, we find a total of 501 institutions with researchers making ten or more extracts (Table 3).   (In addition, in the United States, there are 295 institutions with researchers making 10 or more extracts.)   A surprising number of extracts are made by researchers from countries with no microdata in the IPUMS-I system. The top 10 of these are: Singapore (494 extracts), Belgium (250), Australia (229), Japan (170), Russian Federation (58), Republic of Korea (45), Czech Republic (42), Sweden (41), Hong Kong SAR (40), and New Zealand (40).   On the opposite side of the coin are 14 countries with microdata in the IPUMS-I database but as yet no national researchers use them.   The 14 are: Armenia, Belarus, Ghana, Guinea, Iraq, Jordan, Kyrgyzstan, Mali, Mongolia, Nepal, Peru, Rwanda, Saint Lucia, and Slovenia.   Of course, researchers from these countries——instead of accessing data electronically from the IPUMS-I website—may acquire copies of the integrated microdata on CDs supplied by IPUMS-I to the corresponding National Statistical Office.   We advise NSO partners to register any such users and admonish them to respect the IPUMS-I conditions of use, but these is no requirement to do so.

Interest in comparative research using IPUMS-I extracts is reflected in usage statistics on the mean number of samples requested per extract (Table 3). Since few countries have more than three samples in the database, averages above three suggest research interest in cross-national comparisons, as in Spain (8.3), Austria (4.8), Chile (6.3), Netherlands (7.6), Russian Federation (5.8), etc. The fact that the average is above two, for all but a few countries, indicates that comparative research is pervasive.   Where only one sample is available for a country, it should not be surprising that the average for researchers in that country is also one or nearly so.   In most instances, the 2010 round of censuses will remedy this situation.   In place of one, there will be two samples facilitating comparative research for even the most data-starved countries.

Canada serves as an example of the salience of IPUMS-I research infrastructure for academics and policy makers for a country where access to census microdata is relatively open.   Statistics Canada's Data Liberation Initiative (DLI) dates from 1996 and is widely cited as a model for access to microdata of all types, including census microdata (Goldman 2010).   Canadian users of IPUMS-I rank fifth in terms of number of users (125) and in usage (671 extracts) and fourth in number of institutions (35).   Among Canadian instiutions, the University of Guelph ranks in the IPUMS-I "Top 40" in usage.   Guelph is trailed by seven Canadian universities with 30 or more extracts: British Columbia, Montreal, Queens, Reyerson, Simon Fraser, Toronto, and Western Ontario.   What is surprising—given the success of the DLI and the availability of census samples through Data Research Centers at a dozen or more Canadian Universities—is that 41% of the IPUMS-I extracts by Canadian researchers consist solely of Canadian samples.

The first author queried Canadian users by email and learned that despite the success of the DLI, gaining access to census samples is perceived as tedious and they are troublesome for Canadian researchers.   The metadata, for example, consist of voluminous PDFs, one set per sample, with little guidance as to harmonizing the data from one census to another.   What is equally remarkable about the IPUMS-I statistics is that over half of the extracts by Canadian researchers do not include Canadian samples.   In other words, when Canadian researchers use IPUMS-I extracts in comparative research, more than half do not make use of harmonized Canadian samples.   One explanation is that the Canadian samples (PUFS) are of persons, not households and thus are not readily comparable with 169 of the 185 samples in the IPUMS-I database.   The IPUMS-I "Attach Characteristics" feature for

parents and spouses, for example, cannot be implemented for samples of persons.[1]   Likewise, three of the "Top 33" variables are available only from household samples: MOMLOC, POPLOC and SPLOC.

The lesson to be learned from the Canadian example is that statistical offices disseminating census microdata will gain wider user satisfaction and promote better use by providing access to a complete series of high precision household samples with newly written metadata to facilitate comparative research over time, if not between countries. Economies of scale are achieved, and scarce research resources saved, by integrating both the microdata and metadata, instead of requiring each individual researcher to attempt to harmonize across a series of census samples.   Without integration, researchers will tend to use only one sample.   In the case of Statistics Canada's RDC at the University of Montreal, for example, of 33 successful petitions for access in the academic year 2010/11, only three propose to analyze the complete time series of four censuses.

A second lesson to be learned is that scanned images of old codebooks are no longer sufficient to satisfy user needs.   Nor are microdata files prepared ad hoc over the course of decades with varying sample designs, anonymization procedures, coding schemes and conceptual details.

Integration. IPUMS-I has two rules for integration.   First, retain all significant detail. Second, harmonize every concept and code that appears in two or more censuses.   Note that integration does not mean standardization.   Standardization would require reducing concepts and definitions to their lowest common denominator.   The seeming contradiction of our two simple rules is resolved by the rigorous development of composite, multi-digit coding schemes for each variable. The first digit is for the most general concepts.   The second adds significant detail.   The third and trailing digits, where necessary, contain details that are present in relatively few samples.   If there is no information or additional detail, the digit is coded zero.   For example, marital status (see Figure 2) has only 4 codes for the first digit (at the most general level): 1 - Single, 2 – Married, 3 – Widowed, 4 – Separated/Divorced.   At the second digit, separated is distinguished from divorced.   Married is divided into legal and consensual, and legal marriages may be divided into civil, religious or both. Polygamous unions are also identified by a digit.   The goal is to empower the researcher to make informed decisions about the content and meaning of concepts in the microdata.   With the composite coding scheme, researchers readily understand whether data are suitable for a particular purpose as well as how to recode the data for maximum utility.

To begin the integration process, we first translate to English census forms, instructions to enumerators, codebooks and data dictionaries, if needed.   This step may take a year or two, where there are several censuses and the documentation is particularly voluminous (e.g., Germany and Indonesia).   Second, the MPC integration team applies XML tags to the census documents, associating the variables in the census microdata with the census concepts in the text.   The tagged material in then imported into a database.   Once this step is completed, metadata may be retrieved dynamically for any combination of countries and census years, variable-by-variable.   Initially this tool was developed for the integration team.   Once its utility became apparent, we harnessed the

---

[1]   Another problem with the Canadian PUFS, as a series, is the seemingly erratic suppression of detail. As an example, for the country of birth variable, in most instances, detail is aggregated to the continent, even for countries with fairly large stocks of immigrants, such as China, which is recorded for 1971, 1991 and 2001, but is suppressed for 1981. Hong Kong is recorded for 1991 and 2001.   India is first recorded in 2001.   Greece, Netherlands, and France are recorded for 1971, 1981 and 2001, but suppressed for 1991.   Portugal is suppressed for 1971, and Yugoslavia for 1971 and 1991. The list of countries detailed in all four PUFS is limited to six: Germany, Italy, Poland, Russia/USSR, United Kingdom, and the United States.

dynamic metadata system to the web-site, to permit researchers, whether registered or not, to surf the metadata.

The third step, performed by senior staff, is to reformat the microdata and check for structural anomalies and imperfections (such as two or more heads of households or none, dwellings with no residents or residents with no dwelling, etc.). The fourth step is to confidentialize the microdata (McCaa and Esteve 2005—see wp.5; McCaa, Ruggles and Sobek 2010). Most of the microdata entrusted to the MPC are raw data. Applying uniform confidentiality protocols across countries and census years facilitates comparative research and minimizes infelicities due to variations in confidentiality procedures and errors due to programming mistakes, such as the embarrassment experienced recently by the United States Census Bureau's public use files of the American Community Survey (Alexander, Davern and Stevenson 2010).

Once the microdata are confidentialized, the full integration team, senior staff as well as student research assistants, goes to work, variable-by-variables, searching out unique or undocumented codes, verifying the correspondence of the metadata to the microdata. Issues of comparability of data and census concepts are resolved through discussion and consultation. Ultimately decisions are made, correspondence tables—linking original source codes to integrated composite codes—are finalized, and metadata written to describe nuances that may favor or hinder comparability. For some samples, this process may take three or more years. For most, two years suffice to attain a satisfactory level of integration for most variables and concepts for a country's complete series of censuses. The speed record belongs to Sudan 2008, which was integrated in a mere six months—a record not likely to be surpassed. The IPUMS-I integration process begins with 30-35 samples, for 6-8 countries. When intractable problems are found—usually due to a lack of documentation for codes in the microdata— integration may be postponed for a year or two, until the problems are resolved. Occasionally, serious data editing problems are discovered, which require the expertise of an experienced census data editor. In such cases, with the permission of the corresponding NSO, the microdata are entrusted for resolution, under formal contract, to Dr. Michael J. Levin, contributor to the United Nations Statistics Division *Handbook on Population and Census Editing* (UNSD 2010).

The final step before launch is to generate the IPUMS-I value addeds: sample weights, technical variables (household serial numbers, person numbers, household summary variables), family variables, mother-father-spouse pointer variables, and metadata describing each census and census sample. The annual launch date is June 1.

Future initiatives. The future of census microdata at the MPC is growing brighter as we begin to leverage the power of the microdata beyond the current incarnation of IPUMS-International. Three new initiatives are in various stages of gestation:

1. **SDA** – an online, restricted access tabulator is likely to become operational in 2012. The purpose of the tabulator is to facilitate the experimental research process of registered users. Often researchers wish to ascertain whether a particular research idea is practical, and the tabulator will allow them to explore the data and generate basic tables without having to create and download a data extract. The tabulator is also a useful convenience when a single statistic is all that is desired. Implementing the tabulator will reduce the number of unnecessary extracts, accelerate the research process, and reduce the demand on MPC servers. A version of the SDA is already functioning on the IPUMS-USA site. The tabulator web-page will emphasize that the tabulations are derived from sample data and are not the official population counts of the national statistical offices.

2. **IPUMS-I RDC** – an IPUMS-International Research Data Center for access to full-count and higher density microdata than can be disseminated via the internet, even under

conditions of restricted access. We will develop a secure data enclave at the Minnesota Population Center in 2012 for access to selected data sets.

Our next step is to prototype a system for remote access at secure enclaves at other institutions that agree to enforce the privacy protections necessary for these sensitive data. The system will not deliver the actual data remotely, only the analytic results; and these results will be subject to review by staff at the host institution. This system will be modeled on the best practices for remote-access to confidentialized, higher density census microdata, such as the Australian Bureau of Statistics RADL (Tam, Farley-Larmour and Gare 2009/2010), the Canadian RDC (Goldman 2009/2010), the VML of the United Kingdom (Ritchie 2009/2010) and others. There are two principal differences between these national models and the IPUMS-I RDC. First, researchers, working at "Trusted Centers" anywhere in the world will have access to confidentialized international census microdata instead of microdata for only a single country. Second, both metadata and microdata will be spatially and temporally integrated as closely as possible with the IPUMS-I web-based system. Researchers will be able work inside a Trusted Center to analyze census microdata as they wish, as long as confidentiality is assured. A pilot, using confidentialized, full-count, integrated microdata for two or three countries, is likely to become operational in 2013. Statistical offices interested in considering participation in this initiative should contact the first author of this paper.

3. **TerraPop** – proposes to create a framework for global-scale data on human population characteristics, land use, land cover, and climate change. It will make these data interoperable across time and space, disseminate them to the public and to multiple research communities, and preserve these precious resources for future generations. The TerraPop framework will provide innovative tools for integrating, analyzing, and visualizing data that have spatial and temporal dimensions. More broadly, TerraPop will be a model for the sustainable expansion, maintenance, and improvement of a global data resource.

    Conclusion. The future of population census microdata is bright at IPUMS-International. The project offers a solution for building an integrated metadata and microdata system and managing access to the system on behalf of participating National Statistical Offices as well as academic and policy researchers world-wide. The project demonstrates the substantial economies of scale achievable by working together to build global population census research infrastructures.

In 1999, we proposed to integrate samples for 21 countries, totaling 60-70 censuses. Due to the generous support of national statistical offices and undreamed of economies of scale, 182 samples encompassing 62 countries are now available to researchers—more than double our initial goal. Over the next five years we expect to substantially increase the number of samples as well as extend geographic coverage. Meanwhile a tripling of demand from researchers is easily accommodated with only a modest increase in dissemination costs.

From this foundation, the time is ripe to leverage census microdata with new initiatives—such as the SDA, IPUMS-I RDC and TerraPop—as well as new partnerships with national, regional, and global organizations interested in "opening up the value" inherent in integrated census microdata.

REFERENCES

Alexander, J.T.; Davern, M.; and Stevenson, B.   2010. "Inaccurate Age and Sex Data in the [United States] Census PUMS Files: Evidence and Implications," *Public Opinion Quarterly*, 10 (Aug 10), pp. 1-10.   doi: 10. 1093/poq/nfq033

Conference of European Statisticians.   2007. "Annex 1.23 Case study: Access to anonymized census microdata samples via the IPUMS-International and the Integrated European Census Microdata websites," *Managing Statistical Confidentiality and Microdata Access: Principles and Guidelines on Good Practice*. Geneva: United   Nations   Economic   Commission   for   Europe.   See   online   edition: http://www.unece.org/stats/publications/ pp. 98-104.

Goldman, Gustave. 2010. "From a seed to a forest: Microdata access at Statistics Canada," *Statistical Journal of the IAOS*, 26:75-87.

McCaa, Robert and Albert Esteve.   2005. "IPUMS-Europe: Confidentiality measures for licensing and disseminating restricted access census microdata extracts to academic users," *Joint UNECE/Eurostat Work Session on Statistical Confidentiality*, Geneva, Nov. 9-11.

McCaa, Robert and Albert Esteve.   2009.   "Entrusting census microdata and metadata for timely integration and dissemination via the IPUMS-EurAsia and IECM initiatives, 2010-2014,"   *Census Outputs to Meet User Needs.*   Geneva: United Nations Economic Commission for Europe, Oct. 28-30.

McCaa, Robert, Steven Ruggles and Matthew L. Sobek. 2010. "IPUMS-International statistical disclosure controls: 159 census microdata samples in dissemination, 100+ in preparation,".in J. Domingo-Ferrer and E. Magkos (Eds.): *Privacy in Statistical Data 2010*, LNCS 6344. Springer, Heidelberg, pp.74-84.

McCaa, Robert and Wendy Thomas. 2009, "IPUMS-International: lessons from 10 years of archiving and disseminating census microdata," *International Statistical Institute IPM100.*   Durban, South Africa.

Meier, Ann, Robert McCaa and David Lam. 2011. "Creating statistically literate global citizens: The use of IPUMS-International integrated census microdata in teaching". *Statistical Journal of the IAOS*   forthcoming..

Ritchie, Felix. 2009/2010.   "UK release practices for official microdata". *Statistical Journal of the IAOS* 26:103-111.

Tam, Siu-Ming, Kim Farley-Larmour, and Melissa Gare. 2009/2010. "Supporting research and protecting confidentiality.   ABS microdata access: Current strategies and future directions". *Statistical Journal of the IAOS*   26: 65-74.

United Nations Statistics Division (UNSD).   2010.   *Handbook on Population and Census Editing*. New York: ST/ESA/STAT/SER.F/82.

Figure 1.   a. IPUMS-International project stages of participation: Disseminating (darkest green), integrating (medium green), and negotiating (lightest green).
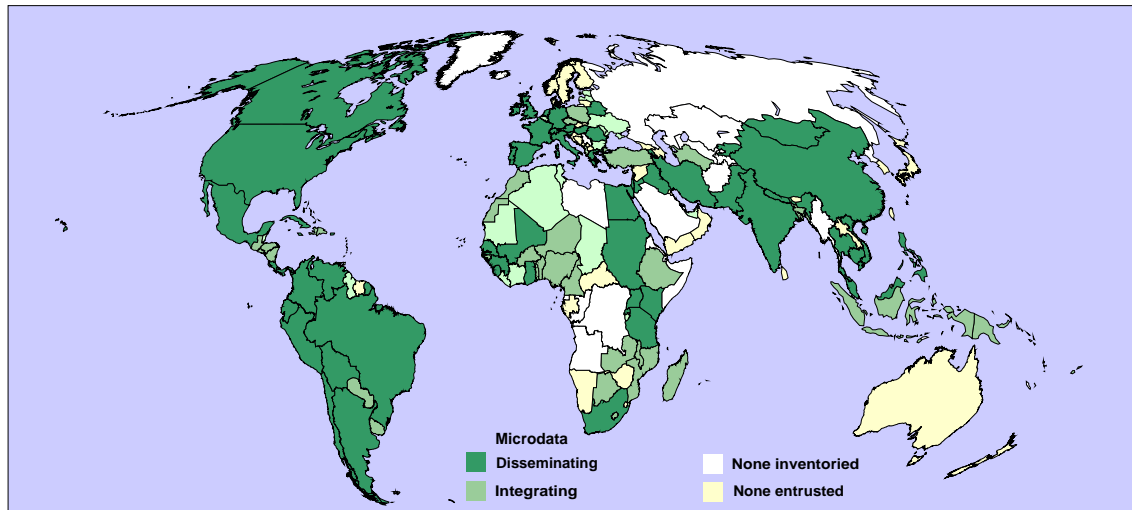
b. Cartogram of IPUMS-International weighted by population size:
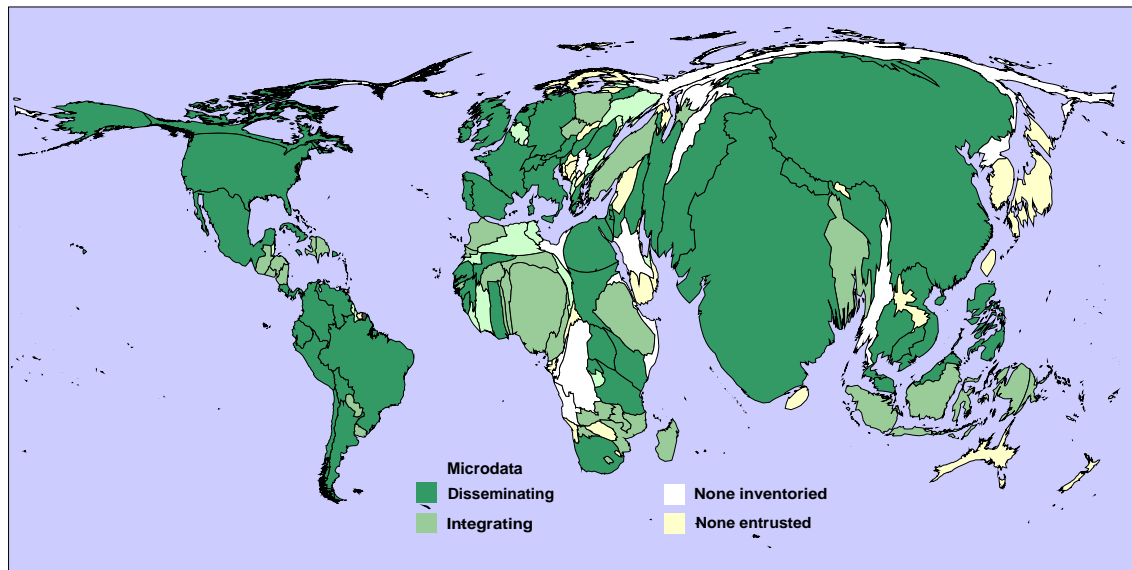
Table 1. IPUMS-International Samples Available to Researchers at https://international.ipums.org: (June 2011)

| Census	%	Persons	Census	%	Persons	Census	%	Persons	Census	%	Persons
1970 **Argentina**	2	466,892	1962 **France**	5	2,320,901	1987* **Malawi**	10	798,669	1980 **Saint Lucia**	10	11,440
1980	10	2,667,714	1968	5	2,487,778	1988*	10	991,393	1991	10	13,405
1991	10	4,286,447	1975	5	2,629,456	2008*	10	1,343,078	1988 **Senegal**	10	700,199
2001	10	3,626,103	1982	5	2,631,713	1970 **Malaysia**	2	175,997	2002	10	994,562
2001 **Armenia**	10	326,560	1990	4.2	2,360,854	1980	2	182,601	2004* **Sierra Leone**	10	494,298
1971 **Austria**	10	749,894	1999	5	2,934,758	1991	2	347,892	2002 **Slovenia**	10	179,632
1981	10	756,556	2006*	33	19,973,287	2000	2	435,300	1996 **South Africa**	10	3,621,164
1991	10	780,512	1970* **Germany**	5	3,094,845	1987 **Mali**	10	785,384	2001	10	3,725,655
2001	10	803,471	1971* DR	25	4,110,749	1998	10	991,330	2007	2	1,047,657
1999 **Belarus**	10	990,706	1981* DR	25	4,278,563	1960 **Mexico**	1.5	502,800	1981 **Spain**	5	2,084,221
1976 **Bolivia**	10	461,699	1987*	5	3,160,224	1970	1	483,405	1991	5	1,931,458
1992	10	642,368	2000 **Ghana**	10	1,894,133	1990	10	8,118,242	2001	5	2,039,274
2001	10	827,692	1971 **Greece**	10	845,483	1995	0.4	332,061	1970 **Switzerland**	5	312,538
1960 **Brazil**	5	3,001,439	1981	10	923,108	2000	10.6	10,099,182	1980	5	317,803
1970	5	4,953,759	1991	10	951,875	2005	10	10,284,550	1990	5	342,797
1980	5	5,870,467	2001	10	1,028,884	1989 **Mongolia**	10	190,631	2000	5	364,086
1991	5.8	8,522,740	1983 **Guinea**	10	457,837	2000	10	243,725	2008* **Sudan**	15	5,609,295
2000	6	10,136,022	1996	10	729,071	2001 **Nepal**	11.4	2,583,245	1988 **Tanzania**	10	2,310,424
1998 **Cambodia**	10	1,141,254	1970 **Hungary**	5	515,119	1960 **Netherlands**	1.2	143,251	2002	10	3,732,735
2008*	10	1,340,121	1980	5	536,007	1971	1.2	159,203	1970 **Thailand**	2	772,169
1971 **Canada**	1	214,019	1990	5	518,240	2001	1.2	189,725	1980	1	388,141
1981	2	486,875	2001	5	510,502	1973 **Pakistan**	2	1,453,332	1990	1	485,100
1991	3	809,654	1983 **India**	0.1	623,494	1981	10	8,433,058	2000	1	604,519
2001	2.5	801,055	1987	0.1	667,848	1998	10	13,102,024	1991 **Uganda**	10	1,548,460
1960 **Chile**	1	88,184	1993	0.1	564,740	1997 **Palestine**	10	259,191	2002	10	2,497,449
1970	10	890,481	1999	0.1	596,688	2007*	10	227,067	1991 **UK**	1	541,894
1982	10	1,133,062	2004	0.1	602,833	1960 **Panama**	5	53,553	2001	3	1,843,525
1992	10	1,335,055	2006* **Iran**	2	1,299,825	1970	10	150,473	1960 **USA**	1	1,799,888
2002	10	1,513,914	1997 **Iraq**	10	1,944,278	1980	10	195,577	1970	1	2,029,666
1982 **China**	1	10,039,191	1971* **Ireland**	10	296,878	1990	10	232,737	1980	5	11,343,120
1990	1	11,835,947	1979*	10	337,686	2000	10	284,081	1990	5	12,501,046
2000		not entrusted	1981*	10	344,291	1993 **Peru**	10	2,206,424	2000	5	14,081,466
1964 **Colombia**	2	349,652	1986*	10	355,020	2007	10	2,745,895	2005	1	2,878,380
1973	10	1,988,831	1991*	10	353,149	1990 **Philippines**	10	6,013,913	1971 **Venezuela**	10	1,158,527
1985	10	2,643,125	1996*	10	365,323	1995	10	6,864,758	1981	10	1,441,266
1993	10	3,213,657	2002*	10	410,688	2000	10	7,417,810	1990	10	1,803,953
2005	10	4,117,607	2006*	10	440,314	1981 **Portugal**	5	492,289	2001	10	2,306,489
1963 **Costa Rica**	6	82,345	1972 **Israel**	10	315,608	1991	5	491,755	1989 **Vietnam**	5	2,626,985
1973	10	186,762	1983	10	403,474	2001	5	517,026	1999	3	2,368,167
1984	10	241,220	1995	10	556,365	1970 **Puerto Rico**	1	27,212	2009*	15	14,177,590
2000	10	381,500	2001 **Italy**	5	2,990,739	1980	5	160,219	**Candidates for June 2012 launch**		
2002 **Cuba**	10	1,118,767	1982* **Jamaica**	10	223,667	1990	5	177,655	El Salvador 2007	10	
1962 **Ecuador**	3	136,443	1991*	10	232,625	2000	5	189,828	Indonesia 1976-10	10	
1974	10	648,678	2001*	10	205,179	2005	1	35,416	Nigeria 2006	5	
1982	10	806,834	2004 **Jordan**	10	510,646	1977 **Romania**	10	1,937,021	Mexico 2010	10	
1990	10	966,234	1989 **Kenya**	5	1,074,098	1992	10	2,238,578	Morocco 1984-2004	10	
2001	10	1,213,725	1999	5	1,407,547	2002	10	2,137,967	Papua New Guinea	10	
1996 **Egypt**	10	5,902,243	1999 **Kyrgyzstan**	10	476,886	1991 **Rwanda**	10	742,918	Turkey 1985-2000	5	
2006*	10	7,282,434	2009	10	forthcoming	2002	10	843,392	Uruguay 1964-1996	10	

*Sample released June 2011

TOTAL: 62 countries — 185 samples — 397,316,462 person records

Table 2.   IPUMS-I Top 40 University and Research Institutions Ranked by Number of Extracts
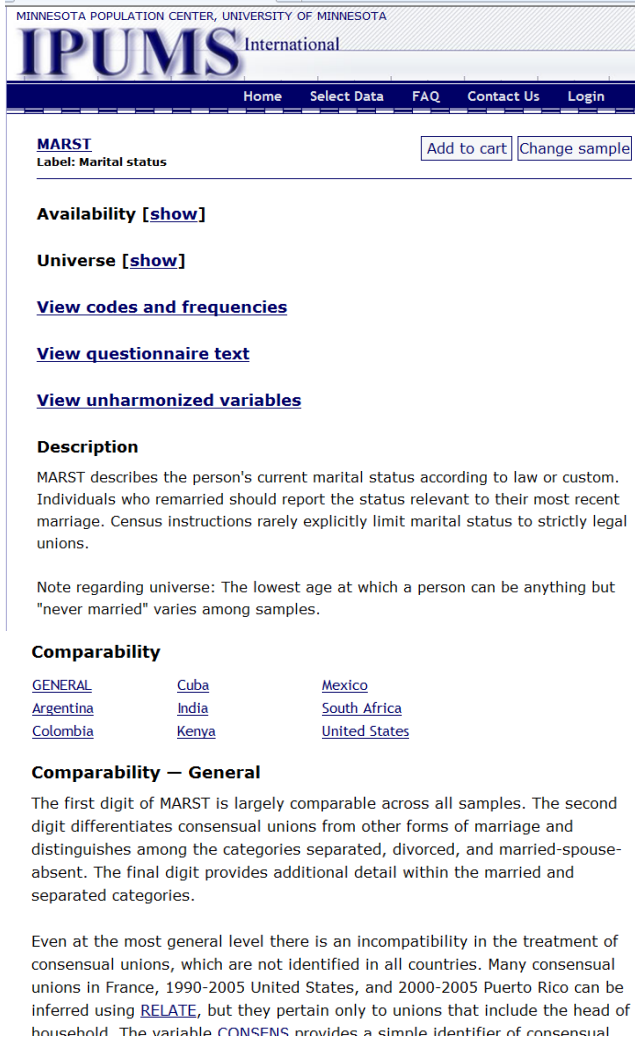
Rank	Institution	N	Rank	Institution	N
1	University of Michigan	742	21	University of North Carolina – Chapel Hill	203
2	Columbia University	701	22	Universite Montesquieu-Bordeaux IV, France	196
3	Universitat de Barcelona, Spain	615	23	University of California - San Diego	189
4	Harvard University	589	24	University of Utah	189
5	Inter - American Development Bank	499	25	World Health Organization, Switzerland	183
6	Arizona State University	495	26	University of Virginia	182
7	National University of Singapore, Singapore	467	27	Michigan State University	178
8	World Bank	408	28	International Institute for Applied Systems Analysis, Austria	165
9	University of California - Berkeley	362	29	University of Sussex, U.K.	158
10	Universidade Federal de Minas Gerais, Brazil	314	30	London School of Economics, U.K.	157
11	University of Chicago	285	31	Dartmouth College	155
12	Universidad del Valle, Colombia	270	32	University of Guelph, Canada	148
13	Institute for Health Metrics & Evaluation	260	33	Institut de Recherche pour le Developpement, France	148
14	Princeton University	237	34	Banco de la Republica, Colombia	145
15	University of Wisconsin - Madison	234	35	Yale University	143
16	Brown University	229	36	University of Tübingen, Germany	143
17	University of Vienna, Austria	229	37	Organization of Economic Cooperation & Development, Fr.	140
18	University of Pittsburgh	227	38	Catholic University Leuven, Brussels	139
19	University of Delaware	213	39	Brigham Young University	138
20	El Colegio de México, México	214	40	University of Queensland, Australia	136

Source: IPUMS-International User Statistics Database, April 18, 2011

Table 3. *Number of Extracts by Researcher's Place of Identity: Samples Extracted and Institution*

Place of Identity	Extracts (N)	Samples Extracted (mean)	Institutions (N)	Place of Identity	Extracts (N)	Samples Extracted (mean)	Institutions (N)
United States	14,669	3.43	295	Ireland	42	2.69	6
France	973	2.95	39	Sweden	41	2.93	8
Spain	972	8.34	23	Hong Kong SAR	40	6.35	5
United Kingdom	961	2.74	41	New Zealand	40	3.23	3
Canada	671	2.35	35	Israel	28	5.04	6
Colombia	627	2.04	16	Pakistan	22	19.59	3
Brazil	598	2.60	22	Puerto Rico	22	1.09	2
Mexico	507	3.33	28	Costa Rica	21	3.62	1
Singapore	494	1.49	4	South Africa	20	4.15	6
Germany	420	3.83	31	Portugal	19	2.32	7
Austria	403	4.77	8	Denmark	18	3.56	1
Italy	377	3.03	27	Senegal	18	2.61	1
Chile	318	6.33	6	Tanzania	13	2.92	2
Argentina	310	3.79	18	Ukraine	13	1.15	1
Switzerland	283	3.92	10	Egypt	10	1.80	3
Belgium	250	2.85	3	Poland	10	1.10	1
Australia	229	2.17	12	Lebanon	9	1.00	1
Netherlands	192	7.58	8	Bosnia and Herzegovina	8	5.63	1
China	184	2.32	25	Cambodia	8	1.00	1
Japan	170	1.68	19	Algeria	7	1.00	1
Kenya	106	1.48	11	Ecuador	7	1.00	1
Greece	92	2.01	7	Norway	7	21.29	2
Russian Federation	58	5.83	5	Syrian Arab Republic	7	2.57	1
Philippines	57	2.18	4	Uruguay	7	10.57	3
Romania	57	6.33	10	Finland	6	2.83	5
Hungary	56	3.68	8	Malaysia	6	2.00	3
India	49	9.06	13	Taiwan	6	3.00	1
Korea Republic of	45	3.87	8	Venezuela	6	2.33	3
Thailand	43	3.30	3	13 Other Places	23	8.33	15
Czech Republic	42	4.69	2	Total	24,699	3.55	835

Figure 2. MARST (Marital Status): 2 screen-shots of metadata for an IPUMS-I integrated variable (note 3 digit composite coding).

a. Description of MARST Variable	b. Case Count View of Detailed MARST Codes – 8 Selected Samples

MINNESOTA POPULATION CENTER, UNIVERSITY OF MINNESOTA

IPUMS International

Home   Select Data   FAQ   Contact Us   Login

MARST
Label: Marital status

[Add to cart] [Change sample]

Availability [show]

Universe [show]

View codes and frequencies

View questionnaire text

View unharmonized variables

Description

MARST describes the person's current marital status according to law or custom. Individuals who remarried should report the status relevant to their most recent marriage. Census instructions rarely explicitly limit marital status to strictly legal unions.

Note regarding universe: The lowest age at which a person can be anything but "never married" varies among samples.

Comparability

GENERAL	Cuba	Mexico
Argentina	India	South Africa
Colombia	Kenya	United States

Comparability — General

The first digit of MARST is largely comparable across all samples. The second digit differentiates consensual unions from other forms of marriage and distinguishes among the categories separated, divorced, and married-spouse-absent. The final digit provides additional detail within the married and separated categories.

Even at the most general level there is an incompatibility in the treatment of consensual unions, which are not identified in all countries. Many consensual unions in France, 1990-2005 United States, and 2000-2005 Puerto Rico can be inferred using RELATE, but they pertain only to unions that include the head of household. The variable CONSENS provides a simple identifier of consensual

Code	Label	argent 2001	colom 2005	cuba 2002	india 2004	kenya 1999	mexico 2000	safrica 2001	usa 2005
000	NIU (not in universe)	·	·	·	·	·	·	·	·
100	SINGLE/NEVER MARRIED	2,168,754	2,188,821	443,373	297,377	903,917	5,533,336	2,436,706	1,147,533
110	Engaged	·	·	·	·	·	·	·	·
200	MARRIED/IN UNION	·	·	·	274,259	·	·	·	1,303,761
210	Married, formally	1,131,429	703,927	312,056	·	·	·	·	·
211	Civil	·	·	·	·	·	878,264	·	·
212	Religious	·	·	·	·	·	208,895	·	·
213	Civil and religious	·	·	·	·	·	2,136,078	·	·
214	Civil or religious	·	·	·	·	·	·	609,273	·
215	Traditional or customary	·	·	·	·	·	·	258,310	·
216	Monogamous	·	·	·	·	389,499	·	·	·
217	Polygamous	·	·	·	·	66,794	·	2,596	·
220	Consensual union	·	750,876	226,692	·	·	765,479	189,157	·
300	SEPARATED, DIVORCED, OR SPOUSE ABSENT	·	·	·	·	·	·	·	·
310	Separated or divorced	·	134,478	·	2,097	·	·	·	·
320	Separated or annulled	·	·	·	·	·	·	·	·
330	Separated	·	·	29,456	·	10,551	177,057	27,036	41,958
331	Separated legally	62,257	·	·	·	·	·	·	·
332	Separated de facto	·	·	·	·	·	·	·	·
333	Separated from marriage	·	·	·	·	·	·	·	·
334	Separated from consensual union	·	·	·	·	·	·	·	·
340	Annulled	·	·	·	·	·	·	·	·
350	Divorced	70,008	·	61,820	·	8,147	56,115	46,462	228,832
360	Married, spouse absent	·	·	·	·	·	·	·	·
400	WIDOWED	193,655	138,708	45,370	29,072	28,639	329,401	156,115	156,296

Code	Label	argent 2001	colom 2005	cuba 2002	india 2004	kenya 1999	mexico 2000	safrica 2001	usa 2005
410	Widowed or divorced	·	·	·	·	·	·	·	·
420	Widowed, divorced, or separated	·	·	·	·	·	·	·	·
999	UNKNOWN	·	89,358	·	28	·	14,557	·	·