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 are catalogued and 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 ninety countries are registered to access confidentialized, integrated microdata without payment and with complete academic freedom—thanks to a uniform licensing agreement endorsed by almost one-hundred 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 sixty-two countries already represented in the database. Samples of 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 world-wide 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 sixty-two countries and totaling 397,316,462 person records are available to researchers free of cost (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 of 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 their generous 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. June 2012, the 2010 round census samples of Indonesia, Mexico, and El Salvador are scheduled for launch—precisely because the data as well as comprehensive documentation were made available without delay.

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 disseminate high precision household samples for approximately 85% of the world's population (80 countries), once the sizeable number of census datasets already entrusted are processed. Thanks to the cooperation of official statistical offices of ninety-eight 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 13 most populous countries yet to embrace the IPUMS-International principles are the Russian Federation, Japan, Congo (DR), Myanmar, Algeria, Afghanistan, Uzbekistan, Korea (RO), Saudi Arabia, Korea (PDR), Yemen, Syria and Australia. 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 are catalogued and are being disseminated world-wide without cost.

In early 2011 IPUMS-I completed a project in cooperation 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 microdata archive. The metadata was structured to be used with the IHSN Microdata Toolkit, developed by the World Bank, which has been introduced in over eighty 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 to direct researchers to IPUMS-International resources.

As part of this project, IPUMS-I has mapped its metadata base and related collection to the DDI standard structure. With this tool, DDI metadata are produced for each extract, customized to each individual request. The DDI can then be rendered as a PDF codebook or be used as input to a web-browser and a growing number of analysis tools that are able to exploit DDI structured documents.

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 poised to strike at the first indication of misuse. 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, in 2010, a second site for Africa (AICMD) 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, which 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 a single, 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 "power users" and novices may be serviced equally well by a single, 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, once registered, the researcher must first log-in by means of a password to place a detailed electronic order ("create an extract"). The next step is to select samples and variables by browsing the corresponding web pages. To review selections, click the data cart. Once the selections are complete, 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. When searching, click "IPUMS-International" to restrict citations to publications using IPUMS-I samples.

As noted above, the usage statistics reveal a surprisingly low average number of variables per extract. 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.

Table 2 reports the number of extracts from the microdata samples of the 55 countries in the database. Mexico, represented by six samples, ranks number one with 9,338 extracts followed by Brazil (6,889), the USA (6,171), and Colombia (4,629). At rank number five, France is the top placed European country in terms of total samples extracted. Six of the next seven rankings are occupied by Latin American countries: Chile (3,492), Argentina (3,152), Ecuador (2,764), Venezuela (2,309), Panama (2,160), and Costa Rica (2,148). The dominance of the Latin American countries in terms of usage is due, on the one hand, to the fact that almost all the official statistical agencies in the region immediately endorsed the idea of integrating the region' s census microdata and granting access through the IPUMS portal and, on the other, to a vigorous tradition of empirical research in Latin American universities. Indeed, two of these, the Federal University of Minas Gerais (Brazil) and the Universidad del Valle (Colombia), rank at the top along with Michigan, Columbia, Harvard, and Berkeley in usage of the samples in the IPUMS database. .

The IPUMS-I "Top 40" institutions in terms of data usage includes many of the world's premier universities and research organizations (see Table 3), scattered across fourteen countries. In 46 countries, we find a total of 501 institutions with researchers making ten or more extracts (Table 4). (In addition, in the United States, there are 295 institutions at this level of activity.) 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 microdata electronically from the IPUMS-I website—may acquire copies of the integrated samples on CDs supplied by the Minnesota Population Center 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 obligation to do so.

Interest in comparative research using IPUMS-I extracts is reflected in the mean number of samples requested per extract (Table 4). 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 of great interest to IPUMS-I researchers. 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 population censuses (Goldman 2010). Canadian users of IPUMS-I

rank fifth in number of users (125) and in usage (671 extracts) and fourth in number of institutions (35). Among Canadian institutions, the University of Guelph ranks in the IPUMS-I "Top 40". 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 troublesome for Canadian researchers. The metadata, for example, consist of voluminous PDFs, one set per sample, with little guidance as to harmonizing the microdata 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 may be that the Canadian samples (PUMFs) 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, is limited to samples of households.[1] Likewise, three of the "Top 33" IPUMS-I 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 broader user satisfaction and promote better use by providing access to a 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. Today's users expect integrated metadata and microdata that are organized to facilitate the research process.

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

[1] Another problem with the Canadian PUFS, as a series, is the seemingly erratic suppression of detail. Take, for example, 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.

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 retain all significant detail in each of the censuses, yet harmonize all concepts. Integration empowers 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 for the research problem at hand.

To begin the integration process, we translate census forms, instructions to enumerators, codebooks and data dictionaries into English, if needed. This step may take a year or two, where there are several censuses and the documentation is particularly voluminous (e.g., Brazil, Germany, Indonesia and Morocco).

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 is 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 to speed the work of the integration team. Once its utility became apparent, we harnessed the dynamic metadata system to the web-site, to permit open access to 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 or nearly so. Names, addresses, and other identifying information are removed, but little else. Working with the "raw" microdata makes it possible to apply uniform confidentiality protocols across countries and census years. Uniform protocols enhances 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). Census agencies that confidentialize data should take heed of this unfortunate episode. Due to a programming mistake age reporting of the elderly was egregiously corrupted in a large fraction of cases in the sample. Researchers could not prove the error until they were able to compare the confidentialized sample against the full-count non-confidentialized microdata available through the Census Bureau's Research Data Center. The brouhaha found its way to the front pages of the New York Times, shortly before the 2010 census got underway. Please be assured that samples confidentialized by the IPUMS-I team are carefully checked for coherence and robustness not only before the microdata are disseminated to researchers but also before the integration work begins.

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, and 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 in comparability. For some samples, this process may take three or more years. For many, 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. Each year, the IPUMS-I final 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 of a specific sample may be postponed for a year or two, until the problems are resolved. If no readily available solution is forthcoming, the entire series of samples for that country is postponed. Sometimes, the launch of the samples for a specific country may be postponed for one, two or even three years or longer while the search for satisfactory original source documentation continues.

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. Finally, the entire group of integrated samples is launched, usually on 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 request an 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 official population counts.
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 are invited to 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. 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, 185 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 to the project—and at no cost to the user.

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* 27(3):145-156.

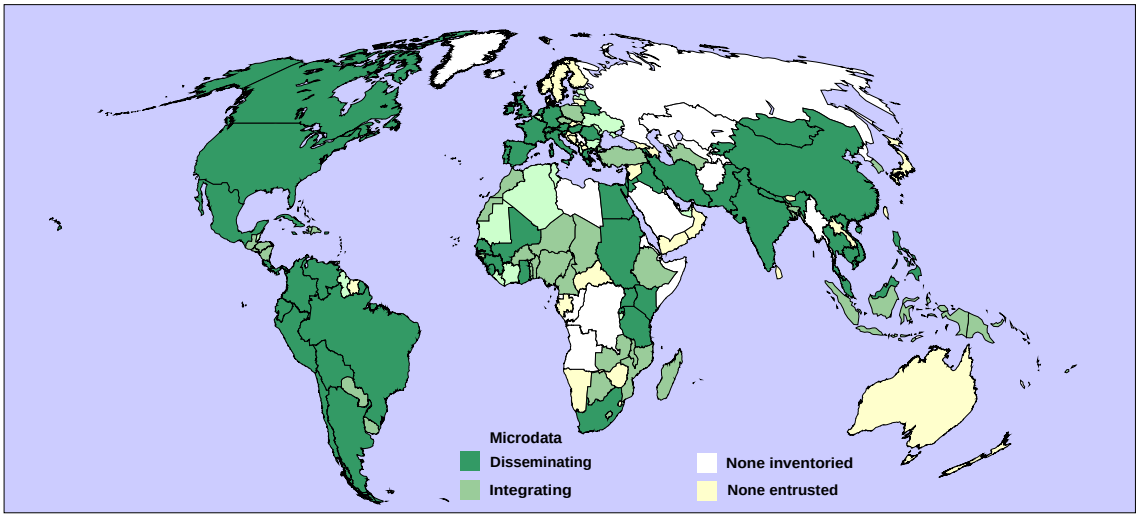
Ritchie, Felix 2010 "UK release practices for official microdata". *Statistical Journal of the IAOS* 26:103-11.

Tam, Siu-Ming, Kim Farley-Larmour, and Melissa Gare. 2009/2010. "Supporting research and protecting

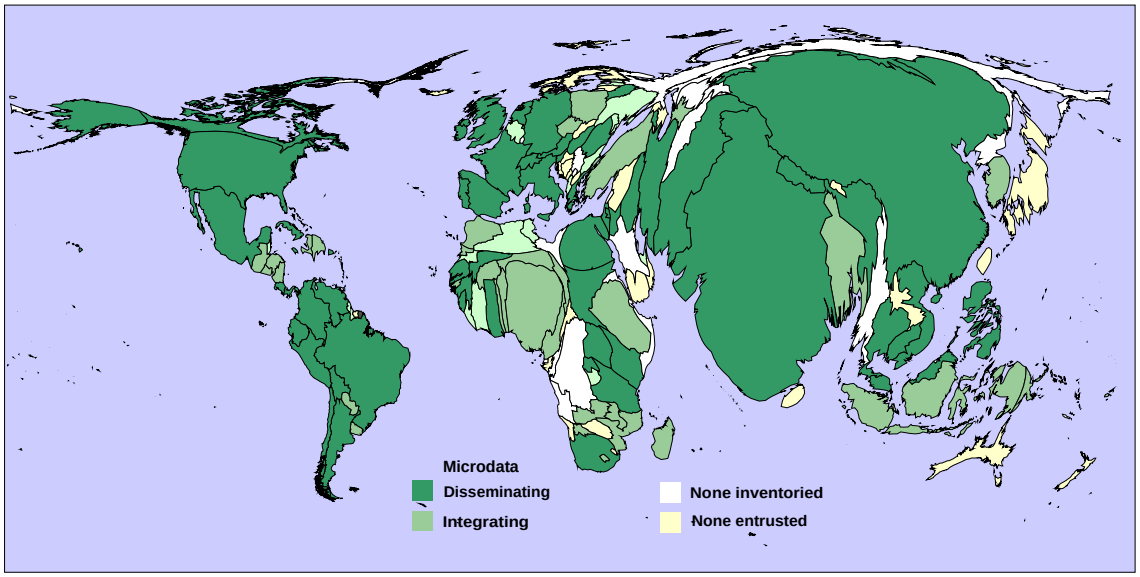
confidentiality. ABS microdata access: Current strategies and future directions". *Statistical Journal of the IAOS* 26: 65-74.

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:



IPUMS-International STS065: The Future of Microdata Access p.12

Table 1. IPUMS-International Samples Available to Researchers at https://international.ipums.org: (June 2011)

Census		%	Persons
1970	**Argentina**	2	466,892
1980		10	2,667,714
1991		10	4,286,447
2001		10	3,626,103
2001	**Armenia**	10	326,560
1971	**Austria**	10	749,894
1981		10	756,556
1991		10	780,512
2001		10	803,471
1999	**Belarus**	10	990,706
1976	**Bolivia**	10	461,699
1992		10	642,368
2001		10	827,692
1960	**Brazil**	5	3,001,439
1970		5	4,953,759
1980		5	5,870,467
1991		5.8	8,522,740
2000		6	10,136,022
1998	**Cambodia**	10	1,141,254
2008*		10	1,340,121
1971	**Canada**	1	214,019
1981		2	486,875
1991		3	809,654
2001		2.5	801,055
1960	**Chile**	1	88,184
1970		10	890,481
1982		10	1,133,062
1992		10	1,335,055
2002		10	1,513,914
1982	**China**	1	10,039,191
1990		1	11,835,947
2000			not entrusted
1964	**Colombia**	2	349,652
1973		10	1,988,831
1985		10	2,643,125
1993		10	3,213,657
2005		10	4,117,607
1963	**Costa Rica**	6	82,345
1973		10	186,762
1984		10	241,220
2000		10	381,500
2002	**Cuba**	10	1,118,767
1962	**Ecuador**	3	136,443
1974		10	648,678
1982		10	806,834
1990		10	966,234
2001		10	1,213,725
1996	**Egypt**	10	5,902,243
2006*****		10	7,282,434

Census		%	Persons
1962	**France**	5	2,320,901
1968		5	2,487,778
1975		5	2,629,456
1982		5	2,631,713
1990		4.2	2,360,854
1999		5	2,934,758
2006*		33	19,973,287
1970*	**Germany**	5	3,094,845
1971*	DR	25	4,110,749
1981*	DR	25	4,278,563
1987*		5	3,160,224
2000	**Ghana**	10	1,894,133
1971	**Greece**	10	845,483
1981		10	923,108
1991		10	951,875
2001		10	1,028,884
1983	**Guinea**	10	457,837
1996		10	729,071
1970	**Hungary**	5	515,119
1980		5	536,007
1990		5	518,240
2001		5	510,502
1983	**India**	0.1	623,494
1987		0.1	667,848
1993		0.1	564,740
1999		0.1	596,688
2004		0.1	602,833
2006*	**Iran**	2	1,299,825
1997	**Iraq**	10	1,944,278
1971*	**Ireland**	10	296,878
1979*		10	337,686
1981*		10	344,291
1986*		10	355,020
1991*		10	353,149
1996*		10	365,323
2002*		10	410,688
2006*		10	440,314
1972	**Israel**	10	315,608
1983		10	403,474
1995		10	556,365
2001	**Italy**	5	2,990,739
1982*****	**Jamaica**	10	223,667
1991*		10	232,625
2001*		10	205,179
2004	**Jordan**	10	510,646
1989	**Kenya**	5	1,074,098
1999		5	1,407,547
1999	**Kyrgyzstan**	10	476,886
2009		10	forthcoming

Census		%	Persons
1987*	**Malawi**	10	798,669
1988*		10	991,393
2008*		10	1,343,078
1970	**Malaysia**	2	175,997
1980		2	182,601
1991		2	347,892
2000		2	435,300
1987	**Mali**	10	785,384
1998		10	991,330
1960	**Mexico**	1.5	502,800
1970		1	483,405
1990		10	8,118,242
1995		0.4	332,061
2000		10.6	10,099,182
2005		10	10,284,550
1989	**Mongolia**	10	190,631
2000		10	243,725
2001	**Nepal**	11.4	2,583,245
1960	**Netherlands**	1.2	143,251
1971		1.2	159,203
2001		1.2	189,725
1973	**Pakistan**	2	1,453,332
1981		10	8,433,058
1998		10	13,102,024
1997	**Palestine**	10	259,191
2007*		10	227,067
1960	**Panama**	5	53,553
1970		10	150,473
1980		10	195,577
1990		10	232,737
2000		10	284,081
1993	**Peru**	10	2,206,424
2007		10	2,745,895
1990	**Philippines**	10	6,013,913
1995		10	6,864,758
2000		10	7,417,810
1981	**Portugal**	5	492,289
1991		5	491,755
2001		5	517,026
1970	**Puerto Rico**	1	27,212
1980		5	160,219
1990		5	177,655
2000		5	189,828
2005		1	35,416
1977	**Romania**	10	1,937,021
1992		10	2,238,578
2002		10	2,137,967
1991	**Rwanda**	10	742,918
2002		10	843,392

Census		%	Persons
1980	**Saint Lucia**	10	11,440
1991		10	13,405
1988	**Senegal**	10	700,199
2002		10	994,562
2004*	**Sierra Leone**	10	494,298
2002	**Slovenia**	10	179,632
1996	**South Africa**	10	3,621,164
2001		10	3,725,655
2007		2	1,047,657
1981	**Spain**	5	2,084,221
1991		5	1,931,458
2001		5	2,039,274
1970	**Switzerland**	5	312,538
1980		5	317,803
1990		5	342,797
2000		5	364,086
2008*	**Sudan (&South)**	15	5,609,295
1988	**Tanzania**	10	2,310,424
2002		10	3,732,735
1970	**Thailand**	2	772,169
1980		1	388,141
1990		1	485,100
2000		1	604,519
1991	**Uganda**	10	1,548,460
2002		10	2,497,449
1991	**UK**	1	541,894
2001		3	1,843,525
1960	**USA**	1	1,799,888
1970		1	2,029,666
1980		5	11,343,120
1990		5	12,501,046
2000		5	14,081,466
2005		1	2,878,380
1971	**Venezuela**	10	1,158,527
1981		10	1,441,266
1990		10	1,803,953
2001		10	2,306,489
1989	**Vietnam**	5	2,626,985
1999		3	2,368,167
2009*		15	14,177,590

Candidates for June 2012 launch

El Salvador 2007	10
Indonesia 1976-10	10
Mexico 2010	10
Morocco 1984-2004	10
Turkey 1985-2000	5
Uruguay 1964-1996	10

***Sample released June 2011**

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

Table 2. IPUMS-I samples integrated and user extracts by country

Rank	Country	Samples	Extracts	Rank	Country	Samples	Extracts
1	Mexico	6	9338	34	Puerto Rico	4	514
2	Brazil	5	6889	35	Ghana	1	510
3	United States	7	6171	36	Switzerland	4	485
4	Colombia	5	4629	37	Cambodia	1	443
5	France	6	3861	38	Belarus	1	437
6	Chile	5	3492	39	Peru	2	399
7	Argentina	4	3152	40	Tanzania	2	384
8	Ecuador	5	2764	41	Italy	1	370
9	Canada	4	2539	42	Mongolia	2	364
10	Venezuela	4	2309	43	Egypt	1	343
11	Panama	5	2160	44	Mali	2	343
12	Costa Rica	4	2148	45	Senegal	2	341
13	Greece	4	2031	46	Palestine	1	334
14	Spain	3	2013	47	Iraq	1	301
15	South Africa	3	1940	48	Pakistan	3	297
16	Kenya	2	1834	49	Armenia	1	263
17	India	4	1700	50	Slovenia	1	242
18	China	2	1635	51	Jordan	1	235
19	Philippines	3	1632	52	Kyrgyz Republic	1	216
20	Austria	4	1608	53	Saint Lucia	2	183
21	Hungary	4	1517	54	Cuba	1	173
22	Vietnam	2	1447	55	Nepal	1	153
23	Portugal	3	1400		**Released June 1, 2011**		
24	Romania	3	1375		Cambodia (2008)	1	
25	Uganda	2	1299		Egypt (2006)	1	
26	Malaysia	4	1275		France (2004-2008)	1	
27	Israel	3	1086		Germany (1970-1987)	4	
28	Rwanda	2	1015		Iran (2006)	1	
29	Bolivia	3	995		Ireland (1971-2006)	8	
30	United Kingdom	2	960		Jamaica (1981-2001)	3	
31	Netherlands	3	810		Malawi (1987-2008)	3	
32	Thailand	4	622		Sudan (2008)	1	
33	Guinea	2	529		Vietnam (2009)	1	

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

Table 3. 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

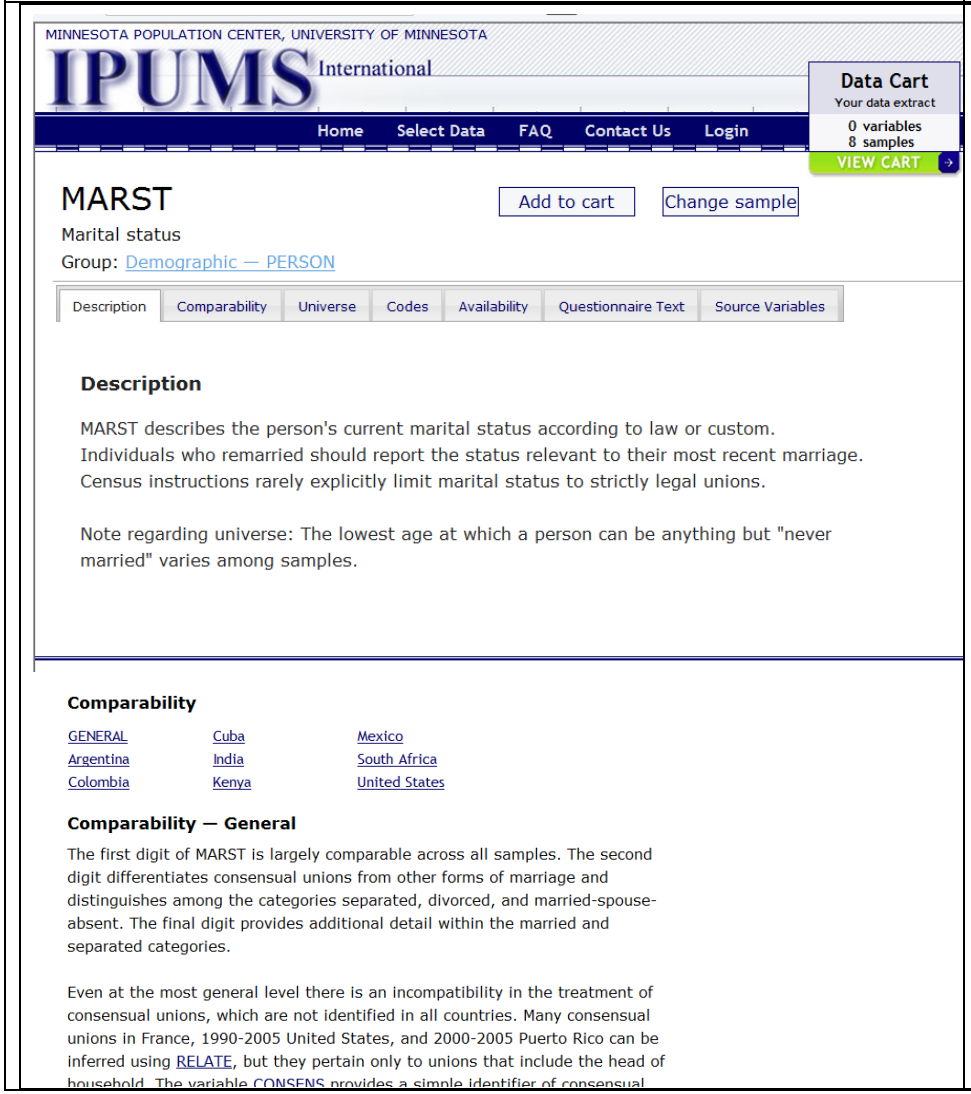
Source: IPUMS-International User Statistics Database, April 18, 2011

Table 4. 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



MINNESOTA POPULATION CENTER, UNIVERSITY OF MINNESOTA

IPUMS International

Data Cart
Your data extract
0 variables
8 samples
VIEW CART

Home | Select Data | FAQ | Contact Us | Login

MARST

Add to cart | Change sample

Marital status

Group: Demographic — PERSON

Description | Comparability | Universe | Codes | Availability | Questionnaire Text | Source 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 | Argentina | Colombia | Cuba | India | Kenya | Mexico | South Africa | 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

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

Code	Label	argent 2001	colom 2005	cuba 2002	india 2004	kenya 1999	m
000	NIU (not in universe)	·	·	·	·	·	
100	SINGLE/NEVER MARRIED	2,168,754	2,188,821	443,373	297,377	903,917	5,
110	Engaged	·	·	·	·	·	
200	MARRIED/IN UNION	·	·	·	274,259	·	
210	Married, formally	1,131,429	703,927	312,056	·	·	
211	Civil	·	·	·	·	·	
212	Religious	·	·	·	·	·	
213	Civil and religious	·	·	·	·	·	2,
214	Civil or religious	·	·	·	·	·	
215	Traditional or customary	·	·	·	·	·	
216	Monogamous	·	·	·	·	389,499	
217	Polygamous	·	·	·	·	66,794	
220	Consensual union	·	750,876	226,692	·	·	
300	SEPARATED, DIVORCED, OR SPOUSE ABSENT	·	·	·	·	·	
310	Separated or divorced	·	134,478	·	2,097	·	
320	Separated or annulled	·	·	·	·	·	
330	Separated	·	·	29,456	·	10,551	
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	
360	Married, spouse absent	·	·	·	·	·	
400	WIDOWED	193,655	138,708	45,370	29,072	28,639	

Code	Label	argent 2001	colom 2005	cuba 2002	india 2004	kenya 1999	m
410	Widowed or divorced	·	·	·	·	·	
420	Widowed, divorced, or separated	·	·	·	·	·	
999	UNKNOWN	·	89,358	·	28	·	