IPUMS-International Disseminates Big Census Microdata

Robert McCaa, Lara Cleveland, and Patricia Kelly Hall
Minnesota Population Center
rmccaa@umn.edu

Abstract: The IPUMS-International initiative (www.ipums.org/international) disseminates integrated census microdata to researchers world-wide free of cost. The IPUMSI database has increased ten-fold over the past decade; with the June 2014 release, the project will encompass 259 anonymized with hundreds of integrated variables samples representing 79 countries and totaling over one-half billion person records. The database is getting bigger in 3 ways: 2010 round census samples (27 done), new partners (25) and a new mode of remote access for higher density samples with greater detail. More than 8,000 users from over 100 countries are currently registered for access. Analytics reveal surprising insights in traffic, usage and publications. Significant challenges remain.

Note: The tables and usage detailspresented in the original abstract submission have been updated for the final poster.

Research for this paper was funded in part by the National Institutes of Child Health and Human Development of the United States, grants HD044154 and 047283 (Latin American, European and Asian census microdata harmonization projects) and by the National Science Foundation of the United States, grant: "International Integrated Microdata Series", SES-0851414, 0433654, and SBR-9908380. The success of the IPUMS-International initiative is due in large part to the generous cooperation and contributions of national and international statistical agencies and research organizations as well as individual official statisticians and academics**.**

Introduction.

The Big Census Data Revolution, foretold a decade ago in Scandinavian Population Studies (McCaa and Ruggles 2002), has arrived, but it is not yet complete. Then, census microdata samples were available for only a handful of countries and trans-border access was difficult for all but a few. Now, from www.ipums.org/international, many decades of census microdata samples for much of the world are readily accessible anywhere, free of cost to researchers and students—regardless of country of birth, residence or citizenship. The website is hosted by the Minnesota Population Center. The IPUMS project is the brainchild of Dr. Steven Ruggles, the Center's founding director. The revolution has sparked much new research. According to a former president of the Population Association of America, students of the Big Census Data Revolution, specifically those with analytical experience using integrated census microdata, enjoy advantages for internships and employment at the World Bank and similar agencies (Meier, Lam and McCaa 2011). Likewise, Dot-Coms beckon as a new jobs frontier opens for savvy Big Data users (Lohr, 2012:B2).

In 1993, the microdata revolution in the United States began with the first release of samples for nine censuses for one country, spanning the period 1880-1990. For academics old enough to remember 7-track computer tape, this was the medium of dissemination with sixty million integrated person records packed on a single reel. Two years later, the first internet website went on-line for Integrated Public Use Microdata Series (IPUMS), and dissemination by tape was quickly forgotten. Today, "IPUMS-USA" disseminates custom-tailored extracts of samples for any of the USA censuses from 1850 to 2010. Each extract is pooled into a single data file, regardless of the number of samples requested. Annually, the USA site is updated with American Community Survey (ACS) samples within a week or two after release by the Census Bureau. The updates include triennial and quinquennial versions of the ACS as well as annual.[1]

[Figure 1 near here.]

In 2002, the IPUMS-International site (https://international.ipums.org) was born, offering pooled extracts of confidentialized, integrated samples for six countries: Colombia, France, Kenya, Mexico, the United States and Vietnam. The ensuing ten years saw a ten-fold increase in the number of countries and samples available to researchers. Usage grew even faster, doubling every two or three years. More than one-half billion integrated microdata records, spanning three-quarters of the world's population, are currently in dissemination.[2] Two continental portals complement the international website with optimized metadata, networking features, and other enhancements:

- The Africa portal, http://ecastats.uneca.org/aicmd/, is hosted by the African Centre for Statistics in Addis Ababa.
- The Europe portal, www.iecm-project.org, is operated by the Center for Demographic Studies in Barcelona.

[1] For a more thorough look at IPUMS-USA and other databases harmonized and distributed by the Minnesota Population Center, see the articles in two special issues of *Historical Methods* (Hall 2011). Data are available from the MPC at https://www.ipums.org/

[2] In addition to the variables harmonized across countries and samples, the IPUMS-International project also makes the original census variables (termed "source variables") for each country and each census available to researchers.

The AICMD is crucial for networking with the large number of African countries, statistical organizations, universities, and researchers over the sprawling, diverse continent. For Europe, the IECM project performs a similar function as well as facilitating linkages with the European Union-funded Data without Boundaries project (www.dwbproject.org).

Global reach. The global reach of the IPUMS-International project is evident in this "Google-Analytics" map of recent traffic to the website (Figure 2). Internet traffic from 125 countries and territories and more than 1,400 cities is depicted on the map. Among cities, Buenos Aires ranks first with more visitors to the site than even Minneapolis, the home of IPUMS-International.

[Figure 2 near here.]

The graphic shows Buenos Aires encompassing much of the Southern Cone of South America. This intense activity was sparked by the release of the integrated sample of the 2010 population census of Argentina, thanks to the close cooperation of the Instituto Nacional de Estadística y Censos (INDEC) of Argentina in quickly completing the data processing of the census and in making the sample available without delay. The big circle for Buenos Aires obscures, but does not entirely conceal, the considerable and broad interest throughout Argentina by researchers and policy makers in accessing the series of integrated Argentine census samples which stretch over five decades, 1970-2010. The Argentine cities of Cordoba, La Plata, Salta, Santa Fe, San Miguel de Tucuman, Bahia Blanca, Mar del Plata, Neuquen, Quilmes, Corrientes, and Resistencia along with Buenos Aires rank among the top 100 cities worldwide for traffic to the IPUMS site. The heavy traffic from South Africa is also quite remarkable.

IPUMS-International Partners

Currently, 100 official statistical agencies participate in IPUMS-International, up from fewer than a dozen ten years ago.[3] Remarkably, once a decision to participate is made, most agencies entrust the country's full series of extant census microdata to the project without undue delay. The list of cooperating countries and territories is as follows (see also the "Partners" link on the IPUMS-International home page):

- Africa (33): Benin, Botswana, Burkina Faso, Cameroon (both the National Institute of Statistics and the Census Bureau), Cape Verde, Central African Republic, Cote d'Ivoire, Egypt, Ethiopia, Ghana, Guinea (Conakry), Guinea-Bissau, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritius, Morocco, Mozambique, Niger, Nigeria (National Bureau of Statistics, but not the National Population Commission which is responsible for the population census), Rwanda, Senegal, Sierra Leone, South Africa, South Sudan, Sudan, Tanzania, Uganda, and Zambia.
- Asia (21): Armenia, Bangladesh, Cambodia, China, India (Ministry of Statistics and Planning Implementation, not the Office of the Census Commission), Indonesia, Iran, Iraq, Israel, Jordan, Korea (Republic of), Kyrgyzstan, Malaysia, Mongolia, Nepal, Pakistan, Palestine, Philippines, Thailand, Turkmenistan, and Vietnam.

[3] See https://international.ipums.org/international/international_partners.shtml for a complete list of national statistical office partners in the IPUMS-International project.

- Europe (20): Austria, Belarus, Bulgaria, Czech Republic, France, Germany, Greece, Hungary, Ireland, Italy, Netherlands, Poland, Portugal, Romania, Slovenia, Spain, Switzerland, Turkey, Ukraine, and the United Kingdom.
- North America (15): Canada, Costa Rica, Cuba, Dominican Republic, El Salvador, Guatemala, Haiti, Honduras, Jamaica, Mexico, Nicaragua, Panama, Puerto Rico, Saint Lucia, and the United States.
- Pacific (2): Fiji Islands and Papua New Guinea.
- South America (10): Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Paraguay, Peru, Uruguay and Venezuela.

Ranked by population size, the nine largest with official statistical agencies yet to participate in the IPUMS-International initiative are: Russia, Japan, Algeria, Saudi Arabia, Korea-DPR, Yemen, Taiwan, Syria and Australia—leaving aside four large countries wholly lacking in census microdata—Congo-DR, Myanmar, Afghanistan and Uzbekistan. As opportunities arise, negotiations continue with statistical agencies in these and other countries not yet inclined to cooperate.

Two-hundred thirty-eight harmonized samples representing 74 countries are currently available to researchers. The microdata are accessible under a uniform license agreement to more than six thousand registered users. Each year, twenty to thirty newly harmonized samples, representing 5-10 countries, are launched. MPC staff and graduate research assistants, led by Dr. Matthew Sobek and Dr. Lara Cleveland, work diligently to complete the time consuming process of documenting and integrating microdata. Samples for the 2010 round of censuses are assigned rush priority to be launched within one year of receipt, where possible.

Over the past ten years, the archival stock of extant census microdata increased by one-third (Table 1). Comparing then and now, 181 sets were thought to be extant for the period 1945-1994 versus 255 today for countries and territories with one million inhabitants or more. For the years prior to 1975, only ten datasets were, in the strict sense of the word, “recovered”—Colombia, Germany (Democratic Republic), Hong Kong, Israel, Liberia, Mongolia, Pakistan, Sudan, Togo, and Turkey. In fact, all microdata from that era require much sleuthing to construct a satisfactorily documented dataset suitable for dissemination. In a couple of cases, our partners keyed microdata from the original enumeration forms. For the older microdata there is often a considerable lag between acquisition by the MPC and its launch by the IPUMS project. Work on the actual microdata cannot begin until the documentation is fairly complete. The integration process consists of two steps. First the metadata must be integrated. Only then can integration of the microdata get underway. With integration completed and checked, the harmonized microdata and metadata are loaded into the IPUMS extract system and dissemination begins to accredited researchers. Notable challenges remain for several historical censuses, specifically: Bangladesh (1981), Italy (1981, 1991), Jordan (1974), Nepal (1991), Spain (1981), Sudan (1973 and 1993), the United Kingdom (1961, 1971 and 1981) and a number of others. Perhaps it is too much to expect that the last three columns of Table 1 will ever become equal — that the number of extant datasets will equal the number held by the MPC and disseminated by IPUMS-International. Nonetheless, with the cooperation of the National Statistical Offices, we expect to narrow the gap.

[Table 1 near here.]

Liberating Trans-border Access to Census Microdata.

Trans-border access to microdata is essential in today's global world, where researchers are highly mobile. Consider, for example, the field of demography, where one-fifth of the membership of the International Union for the Scientific Study of Population (IUSSP) resides outside the country of birth. Of the 506 members resident in the USA, thirty percent were born elsewhere. One-third of IUSSP demographers born in China do not presently reside there. For German and Dutch-born IUSSP members, the fraction of expatriates is even higher.[4] For many demographers—and many statisticians, economists, and social scientists in general—trans-border access is essential if analysts are to research census microdata of their country of birth as well as engage in comparative, cross-national research. All microdata in the IPUMS-International system are accessible to bona fide researchers world-wide on identical terms. Table 2 shows the distribution by geographic region of the harmonized population data available to researchers through the IPUMS-International partnership.

[Table 2 near here.]

As recently as ten years ago, many official statistical agencies were reluctant to grant access to census microdata even for their own national researchers, much less non-nationals. Today, official statisticians who deny access find themselves on the defensive. Fortunately, many now understand the importance of international microdata access and work to find solutions to facilitate dissemination, including by third parties such as IPUMS. Others seek favorable administrative rulings and some even draft legislation to modernize their statistical charters to facilitate international dissemination of census microdata.

Nonetheless, there are agencies that continue to deny access or erect barriers with archaic rules. In negotiating agreements, I do not inquire as to the reasons for denial, but reasons—justifications, rationalizations or excuses—are often volunteered. They are sometimes clothed in thin, ill-fitting garments of law (to which one might reply: "amend the law"), privacy (apply disclosure controls), popular opposition (publicize the benefits), custom (chat with the younger generation), inertia (let's just do it), secrecy (risk your job), public order (is there a single instance of a riot ensuing from knowledge of microdata?), etc. Fortunately the vast majority of statistical agencies understand the benefits to be gained by facilitating international access.

Microdata disseminated by IPUMS-International are governed by uniform legal and administrative protections and are subjected to strong technical statistical disclosure controls. This approach provides greater protections for the group of statistical agencies as a whole than for any single office that chooses to "go it alone" (Cleveland, McCaa, Ruggles and Sobek 2012). To maximize effectiveness, disclosure controls for access to census microdata must be legal, administrative and statistical (Thorogood 1999). Otherwise utility is sacrificed on the altar of risk. Access to the IPUMS-International microdata is restricted—despite the "P" (Public) in IPUMS—governed, on the one hand, by the letter of understanding endorsed by the University

[4] Statistics provided to the authors by the Secretariat of the International Union for the Scientific Study of Population, September 14, 2011.

and the National Statistical Authority, and, on the other by the license agreement between the University, the researcher, and the researcher's institution. The letter of understanding grants to the University rights to disseminate microdata extracts electronically for teaching and research. According to the authorization procedures stated in the agreement, microdata may not be used for commercial purposes. Strict confidentiality of persons, households and other entities must be maintained. Alleging that a person or other entity has been identified is prohibited. Users must also guard against access to the microdata by unauthorized individuals. The usage license is for one year and may be renewed. The University of Minnesota is the enforcer of the license agreement.

Non Disclosure Protocols.

With respect to technical statistical controls, the first and most important privacy protection is the suppression of names and low-level geographic details. The second is the use of sub-sampling to suppress records. For most samples, 90% of all person records are suppressed, leaving only 10% for research. What this means is that all the values in the records outside the sample are excluded. Third, each statistical authority balances the risk-utility trade-off by instructing the IPUMS project as to the minimum thresholds for identifiable social categories and geographical units for the most recent census. For social categories, population minima are often set at 250 individuals, but in some cases the number is 2,500 or even higher. The geographical threshold is commonly set at 20,000 inhabitants. Some agencies set the floor as high as 100,000 (United States) or in the most extreme case (Netherlands) all administrative geography is suppressed. We are gratified that in several instances our statistical agency partners have reconsidered earlier, overly strict decisions, to approve higher precision samples (Mexico 1990 increased from one to ten percent) and greater detail. In the case of Colombia, the geographical threshold, initially set at 100,000, was reduced to 20,000 after Colombian researchers vigorously lobbied the national statistical agency, DANE. DANE not only reduced the threshold, but also harmonized the geographical codes so that all the census microdata samples for Colombia could be disseminated with a single set of integrated geographical identifiers, in harmony with national practice. When the sample for the 2005 census of Colombia became available, applying uniform standards for confidentializing and harmonizing geographical codes for the complete series of censuses, 1964-2005, was easily accomplished.

Additional technical controls include: top/bottom-coding of continuous variables, global recoding of categorical variables, suppressing digits of hierarchical variables (occupation, industry, geography), suppressing sensitive variables entirely (e.g., "tribe" in Kenyan census microdata), etc. Additional statistical disclosure protections are provided by randomly ordering the records and actually swapping the geographical identifiers of an undisclosed number of households. Swapping corrupts the geographical integrity of the microdata to a small degree, but doing so provides a powerful argument that no one can allege with absolute certainty that an individual or household has been identified. Weight variables and expansion factors are usually not an issue because most of the samples are implicitly stratified so that all records carry an identical weight, such as 10 for a ten percent sample. These and many other decisions are made in consultation with each national statistical authority. Often responsibility for implementing statistical disclosure protections is entrusted to IPUMS project managers.

IPUMS privacy protocols offer strong disclosure control protections at modest cost in terms of loss of statistically useful information. They also protect against the introduction of biases or bugs (errors) into the microdata. Microdata corruption is a grave concern of researchers as more statistical agencies assume the role of imposing confidentiality protections (see Reiter 2011; Cleveland et. al. 2012, and Alexander, Davern and Stevenson 2010, regarding the inadvertent corruption of the American Community Survey).

Trans-Border Dissemination of Census Microdata.

IPUMS disseminates pooled extracts containing many samples in a single dataset, custom-tailored to the precise research needs of the user. This contrasts with the practices of most statistical offices where microdata are disseminated, if at all, as a single dataset containing all variables and all person records in the sample for each census. The common practice has been for every researcher to receive exactly the same dataset. The circulation of a single dataset tempts the unauthorized to seek an illicit copy. With IPUMS-International each extract is unique. Therefore each researcher is nudged into cooperating to guard the data from unauthorized persons. Given the massive size of the IPUMS-International database, disseminating the full set of variables and unvarying size of samples is impractical. Most importantly, IPUMS disseminates pooled microdata with multiple samples and a varying selection of variables for each extract request. This is possible because both microdata and metadata are integrated for all censuses and all countries. 37% of extracts in 2011 requested more than one sample. The average number of variables extracted was 35, including six technical variables that are mandatory with each extract.

With IPUMS no two extracts are alike. Each extract is custom-tailored. The researcher places an order, selecting:

- country (or countries)
- census year(s)
- variables (age, sex, educational attainment, etc.)
- sub-populations (e.g., female heads of households aged less than twenty five years along with all other co-resident persons in the selected household)
- and sample density (either as a percent or number of cases).

The IPUMS extract engine fulfils the request by generating a dataset containing only the requested microdata and the corresponding set of DDI (Document Data Initiative) compatible metadata including a codebook suitable for constructing a system file in SPSS, SAS or STATA. Copies of original source metadata are available from the website. Most importantly the integrated metadata are always readily available in interactive form from the website.

In 2005, at the UN-ECE Expert Group Meeting on Statistical Data Confidentiality, we summarized the IPUMS-International data dissemination procedure as follows (McCaa and Esteve 2005):

> When the extract is ready (usually in a matter of minutes), the researcher is notified by email that the data should be retrieved within 72 hours. A link is provided to a password-protected site for downloading the specific extract. The data are encrypted during transmission using 128-bit SSL (Secure Sockets Layer) encryption standard, matching the level used by the banking and other industries where security and confidentiality are essential. The researcher may then securely download the extract, decompress it and

proceed with the analysis using the supplied integrated metadata consisting of variable names and labels.

This method of dissemination has weathered the test of time, and indeed as usage soars, the rapid acceleration of internet transmission speeds has validated the IPUMS approach. Nonetheless, we continue to seek more secure and efficient ways to facilitate researcher access.

In 2011, 8,048 extracts were made from the IPUMS-International website, totaling 40,142 samples and 281,640 variables. The average number of extracts per country was almost 150 samples for the 55 countries represented in the database for the full year (Table 2). Nonetheless, usage by country varied greatly. The smallest number of extracts, 127, was registered for the 1997 census of Palestine. The greatest, 712, was registered for the sample from the 2000 census of Brazil.

[Table 2 "Regional Distribution of Data" near here.]

Brazil, Mexico and Colombia predominate in usage not only because their samples offer many variables and a long chronological series covering a half century of dramatic demographic transformations, but also due to the fact that many Latin American emigrants reside in the United States or Spain and thus it is possible to analyze these populations in a single integrated database, regardless of where the researcher resides. In addition, with the exception of the oldest samples, all the Latin American data, as well as those for the United States and Spain, are high precision, household samples with richly detailed, extensive information on migration, economic, social and demographic variables for both individuals and households.

For the year 2011, 1,011 researchers qualified for access to the IPUMS-International database, representing 98 countries. The IPUMS "Top 33" institutions in terms of data usage represents fourteen countries and territories and include some of the world's premier universities and research organizations (Table 3).

[Table 3 "Top 33" near here.]

Integrated, Pooled Microdata and Metadata.

The principal benefit of IPUMS to researchers and National Statistics Offices alike is the integration of several decades of microdata samples for each country—typically beginning with the earliest census for which microdata exist or are recoverable. While some NSOs have provided a sample for the most recent census, few re-examine earlier censuses to produce a cross-walk table to harmonize successive samples. Nor is much effort given to drafting new documentation to facilitate comparative analysis of two or more censuses. Most statistical offices are severely under-staffed and face significant financial and human resource constraints. Where microdata dissemination is considered at all, the practice has been for the office to construct a census sample and a data dictionary. Five or ten years later, once data processing is completed, the process is repeated. Little guidance is offered on how to compare microdata from successive censuses. The sample for each census remains unaltered as a single file, leaving each individual researcher the task of attempting to construct a series over time. Most researchers faced with such a daunting undertaking simply resort to analyzing the most recent sample and ignoring earlier data.

With IPUMS-International, researchers are empowered to analyze multiple census years and even multiple countries as a single dataset, facilitating comparative analysis over time and space. High-precision census samples are integrated, variable-by-variable, using a composite coding system (Esteve and Sobek, 2003). Samples are integrated both chronologically and cross-nationally. Integrated metadata are constructed from the meticulous study of comprehensive original source documentation accompanied by extensive analysis of the microdata. Thousands of hours are devoted to analyze, discuss, debate, test and re-test until the microdata integration is validated for dissemination to researchers. The process is repeated with each annual launch of additional census samples into the IPUMS database.

The basic goal of the harmonization effort is to simplify the use of the data while losing no meaningful information. This is a challenging task because to make data simple for comparative analysis across time and space, it is necessary to develop comparable coding schemes. Microdata are integrated so that identical concepts (variables, categories) have identical codes. To avoid the loss of important information for those samples that have even more detail, a composite coding strategy is used to retain all original detail, and at the same time provide comparable codes across samples. With composite codes, researchers may easily compare across time and space, yet nuances in meaning are readily discernible. The first digit, called the "general code," provides information that is available across all samples (the lowest common denominator data). The next one or two digits provides additional information available in a substantial subset of the samples. Trailing digits provide detail that is only rarely available. Where information is not available for a particular sample, a zero place-holder is assigned to that digit.

As an example of this method of integrating variables, consider the concept "educational attainment," the single most widely used variable in the IPUMS-International database. Most census microdata with information on this measure indicate whether the respondent completed primary, secondary or higher schooling or no schooling at all. Thus the first digit of the IPUMS-International composite code consists of four categories (1-4), plus codes for missing data (9) and "not in universe" (0—for children too young to attend or others to whom the question was not addressed). Many census samples contain further information indicating, for example, those who attended primary, secondary or even tertiary schooling, but did not complete the course of study. The second digit captures this information. The third digit distinguishes between technical and general or other tracks common to two or more countries. Successful international integration must document such distinctions so that researchers may readily be informed of these and thousands of other details.

Table 4 illustrates the general and detailed coding schemes for the educational attainment variable for 16 countries (represented by its two-digit ISO 3166 code) and census samples (represented by a two-digit year code with century omitted). As the upper section of the table shows, all samples have each of the four general codes: less than primary completed, and primary, secondary and tertiary completed. In the lower section of the table, the array of detailed codes displays the considerable variability from country-to-country in the level of specificity regarding the various tracks of schooling completed. In addition to these codes, the IPUMS metadata offers general descriptions, comparability discussions, statements of universe, availability of concepts, detailed wording of the original texts and links to the source documents in the official language and English translation. The goal is to facilitate informed analysis of the microdata by providing as much essential information as possible—all readily accessible from the website by means of a few clicks.

[Table 4 "Educational Attainment Codes" near here.]

IPUMS-International Value-Added Services to Researchers.

The IPUMS-International project provides an enormous service to the research community through the cross-temporal and geospatial integration of population data. There are also additional features of IPUMS-International data that add significant value to the raw data for social research. These include extensive integrated metadata--especially documentation and guidance on the variability in sample design; a data tabulator, within-household relationship pointer variables, and GIS boundary files.

Sample documentation and guidance. The microdata samples in IPUMS-International employ a variety of sample designs. All IPUMS samples contain individual level data, most are clustered by household, many are stratified, and some are differentially weighted. Detailed sample descriptions (https://international.ipums.org/international/samples.shtml) and guidance for variance estimation (https://international.ipums.org/international/variance_estimation.shtml) and provided in the extensive documentation on the website and in Cleveland, Davern and Ruggles (2011).

Tabulator. Quick tabulations can now be made with the IPUMS International Online Data Analysis System. The IPUMS online analysis system uses high-speed tabulation software developed at UC-Berkeley's Computer-assisted Survey Methods Program. Researchers registered with IPUMS International may specify samples and variables of interest and get quick calculations output to their computer screen or mobile device.

Pointers, IPUMS data also include powerful constructed variables that aid researchers in utilizing information about household structure implicit in the census data samples. These variables, referred to as "pointers," include a consistent, versatile, and reliable set of constructed variables that describe a variety of family interrelationships among individuals within the same household. Researchers can use them to easily link characteristics of one family member to another - spouses to spouses, children to either or both parents, and so on - thereby speeding up analyses of family structures and characteristics. Methodology for the construction of IPUMS International pointer variables can be found in the working paper by Sobek and Kennedy (2009).

GIS boundary files. As geospatial measurement techniques have advanced, so has user demand for additional geographic information and tools for utilizing spatial data. We have recently enhanced documentation and harmonization of spatial information contained in the household geography of the census records. IPUMS-International has added spatiotemporally harmonized geographic variables and accompanying boundary files (shapefiles) to facilitate national and international data mapping. Users can create maps with IPUMS-International data using a statistical software program and ArcGIS (a GIS mapping software).

Research results: Bibliography

The ready availability of census microdata is beginning to bear fruit. Researchers are required to file citations of research results in the on-line bibliography, as a condition of the license agreement (http://bibliography.ipums.org/quick_submissions/new). To encourage timely compliance, we have begun to delay requests for renewal by those lacking recent citations in the website bibliography. Although the bibliography will always remain incomplete with the lag time between completion of research and publication, we can now take stock of a decade of cited works. Readers are invited to peruse the on-line bibliography to make a personal assessment. From the bibliography webpage, clicking the "IPUMS-International" project box filters out citations from other MPC managed projects. However be forewarned that some citations are incorrectly tagged. For this overview, I excluded about one-fifth of the hits and compiled a carefully targeted set of 450 citations. Among these are a half-dozen books, a dozen World Bank studies, two dozen dissertations and more than 100 journal articles. Most of the major demographic journals are represented. Population and Development Review ranks at the top of the list with eight citations. Demography published one of the most widely cited articles (Van Hook and Glick, 2007) with 42 citations (Google Scholar). Van Hook and Glick coupled the small, highly specialized Survey of Income and Program Participation with census microdata for the United States and Mexico to better interpret the results of immigration on household structure. In reading the abstracts of the citations, one quickly learns that, like Van Hook and Glick, many researchers exploit a variety of data sources, and do not rely solely on census microdata.

In terms of geography, over half of the citations focus on a mere six countries: Mexico, Brazil, South Africa, Colombia, Chile and China. Expanding the focus to one-third of the countries in the database raises the proportion to 90%. The additional 16 countries are: France, Argentina, India, Kenya, Canada, Spain, Ecuador, Uganda, Vietnam, Romania, Rwanda, Costa Rica, Germany, Ghana, Venezuela, and Greece. Not surprisingly, these rankings are strongly correlated with the length of time the microdata for a specific country has been available from the IPUMS-International website. As the database matures, and particularly as 2010 census round samples become accessible, a much greater geographical diversity in published results will likely emerge.

The distribution of citations by subject matter is probably already well established based on the first decade of publications. Among the thirteen broad classifications offered by the on-line bibliography, three account for almost half the citations: labor force and occupational structure, migration and immigration, and family and marriage. A second group of three—education, methodology and data collection, and fertility and mortality—swells the total to three-quarters. A group of five subjects is tied at roughly four percentage points each: education, methodology and data collection, fertility and mortality, gender, and aging and retirement. Housing and segregation studies account for less than 3% and crime and deviance 0.1%. Note that more than half of the citations are listed with more than one subject.

Conclusion.

When we began over a decade ago, we dreamed of integrating census microdata for perhaps a couple of dozen countries in ten years. Thanks to the generous cooperation of National Statistical Offices and undreamed of technological innovations, the number of countries approaches six dozen, and integration work continues. The number of users and the amount of usage also far exceed expectations. For the second decade, we dream of doubling the number of users and doubling again the number of samples. High-precision samples for the 2010 round of censuses will be crucial to our success as well as the integration of samples for the twenty-five partners already in line. The Remote Access Data Enclave will complement the IPUMS-International database with higher density samples—in some instances full count microdata—and variables with greater detail than can be disseminated in the form of individual person records.

Bibliography.

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

Cleveland, L.; McCaa, R.; Ruggles, S.; and Sobek, M. 2012. "When Excessive Perturbation Goes Wrong and Why IPUMS-International Relies Instead on Sampling, Suppression, Swapping, and Other Minimally Harmful Methods to Protect Privacy of Census Microdata, " In J. Domingo-Ferrer and I. Tinnirello (Eds.): *Privacy in Statistical Data 2012, LNCS vol 7556.*

Cleveland, L., Davern, M. and Ruggles, S. 2011. "Drawing Statistical Inferences from International Census Data." *IPUMS International Working Paper.*

Cogburn, D. L. (2003). HCI [Human Computer Interaction] in the so-called developing world: what's in it for everyone, *Interactions, 10*(2), 80-87, New York: ACM Press.

Esteve, Albert, Matthew Sobek. 2003. "Challenges and methods of international census harmonization ." *Historical Methods* 36: 66-79

Hall, Patricia Kelly (ed.) 2011. "Big Data: Large-Scale Historical Infrastructure from the Minnesota Population Center," *Historical Methods Special Issue, Part 1 (Vol. 44, No. 1), and Part 2 (Vol. 44, No. 2)*

Lohr, Steve. 2012. "New U.S. Research Will Aim at Flood of Digital Data," *New York Times*, March 29, p. B2.

McCaa, Robert and Albert Esteve. 2005. "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, Steven Ruggles, Michael Davern, Tami Swenson, Krishna Mohan Palipudi. 2006. "IPUMS-International High Precision Population Census Microdata Samples: Balancing the Privacy-Quality Tradeoff by Means of Restricted Access Extracts." Pp. 375-382 in *Privacy in Statistical Databases*. New York: Springer

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

Meier, Ann, Robert McCaa, David Lam. 2011. "Creating statistically literate global citizens: The use of IPUMS-International integrated census microdata in teaching," *Statistical Journal of the IAOS* 27: 145-156

Sobek, Matthew, Lara Cleveland, Sarah Flood, Patricia Kelly Hall, Miriam King, Steve Ruggles, Matthew Schroeder. 2011. "Big Data: Large-Scale Historical Infrastructure from the Minnesota Population Center." *Historical Methods* 44: 61-68

Sobek, Matthew and Sheela Kennedy. 2009. "The Development of Family Interrelationship Variables for International Census Data." *MPC Working Paper Series 2009-02.* (http://www.pop.umn.edu/sites/www.pop.umn.edu/files/Working%20Paper%202009-02.pdf)

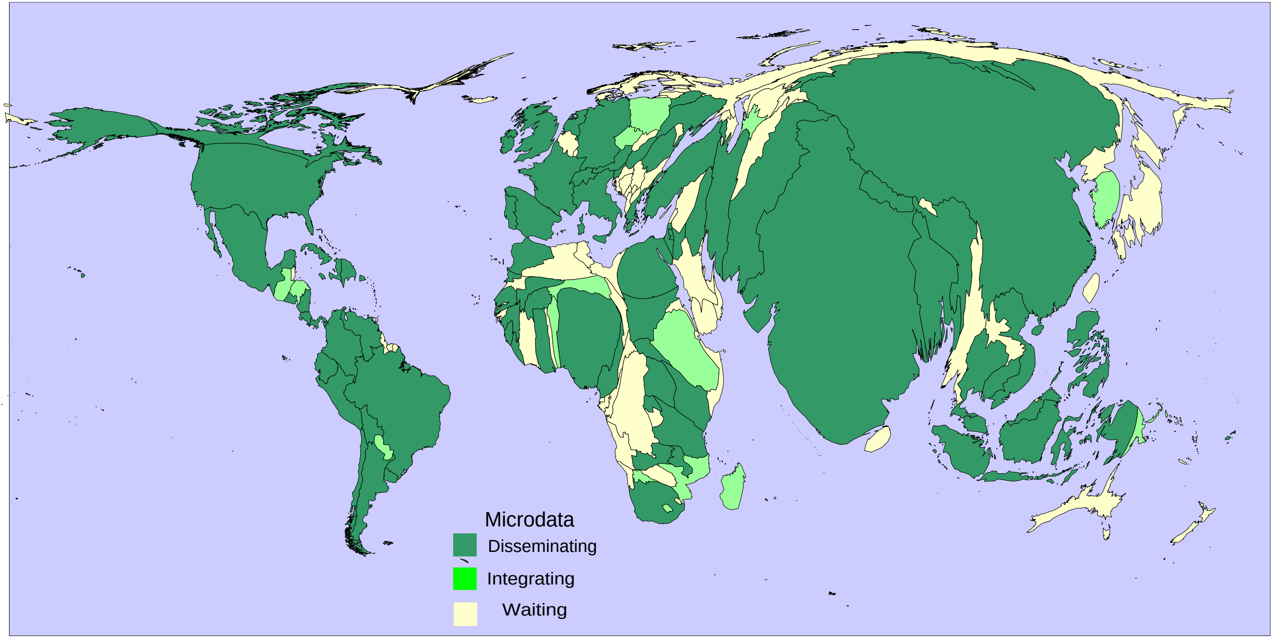
Thorogood, David. 1999. "Statistical Confidentiality at the European Level," *Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Thessaloniki, March.

Van Hook, Jennifer, Jennifer Glick. 2007. "Immigration and Living Arrangements: Moving Beyond Economic Need versus Acculturation." *Demography* 44: 225-249.

Figure 1: IPUMS-International Disseminates Big Census Data

561,622,889 person records • 259 samples • 79 countries

(The area of each country shown is weighted by that country's proportion of the world's population in 2014.)



The IPUMS-International disseminates population data for 82 percent of the world's population, and data from 79 countries and 259 census samples. We are currently integrating data from 18 countries (4 percent of the world's population) with data for an additional 14 percent of population scheduled for future integration. However, 13 of the world's largest countries have not yet endorsed IPUMS-International data-sharing protocols; these include Russia, Japan, Congo, Democratic Republic, Myanmar, Algeria, Afghanistan, Uzbekistan, Taiwan, Korea DPR, Saudi Arabia, Australia, Sri Lanka and Yemen.

Table 1. Growth in Available Population Census Microdata over a Dozen Years (158 Countries/Territories: 1 million plus population in 2014)

Population censuses became universal in the late 20th century; trans-border dissemination of census microdata is becoming universal in the first decades of the 21st century.

	Number	% of World	Country Inventory			
Census	Conducting	Population			Extant Microdata	
Round	a Census	Enumerated	2002	2014	MPC	IPUMS-Int.
1945-54	86	79	2	2	1	0
1955-64	116	87	27	24	22	12
1965-74	129	73	44	51	44	32
1975-84	138	96	54	75	62	41
1985-94	137	96	54	103	91	55
1995-04	129	85	-	122	85	73
2005-14	149*	91*	-	129	53	46
Total	884	-	181	511	358	259

Notes: (1.) China: most recent sample integrated: 1990. Awaiting 2000 and 2010; (2.) India: National Sample Survey Organization Schedule 10 samples. (3.) Nigeria: General Household Surveys. (4.) For the 1960 round, two datasets--Canada and Philippines--thought to exist in 2002 are not yet usable, and for a third--Austria--the only known copy was inadvertently destroyed. (5) *For the 2005-14 round the number of censuses is provisional (our computations from: http://unstats.un.org/unsd/demographic/sources/census/

Table headings: "Extant" - number of countries/territories with confirmed census microdata in existence.
"MPC" - number entrusting microdata (sample or full-count) to Minnesota Population Center.
"IPUMS" - number of samples currently integrated & disseminated from www.ipums.org/international.
Sources: McCaa and Ruggles (2002), Table 1; and "IPUMS-International microdata inventory":
http://www.hist.umn.edu/~rmccaa/IPUMSI/census_microdata_inventory.htm

Table 2. Number of Person Records Available in IPUMS-International by Geographical Region

	Countries	Samples	Person Records
	N	N	N
Africa			
Eastern	6	17	25,929,627
Middle	1	1	2,669,570
Northern	4	7	23,126,356
Southern	1	1	8394476
Western	8	20	15,273,334
Asia			
Central	1	1	1605362
Eastern	2	4	22,309,494
South-Eastern	6	26	101,272,939
Southern	5	13	60,155,826
Western	6	11	13,406,216
Caribbbean	6	20	6,657,049
Central America	5	22	46,133,884
Europe			
Eastern	4	9	14,273,428
Northern	2	11	5,763,121
Southern	5	12	14,790,410
Western	5	22	54,882,071
North America	2	11	50,006,861
Oceania	1	5	338,656
South America	9	42	94,912,843

Source: https://international.ipums.org/international-ction/variables/REGIONW#codes_section

Table 3. IPUMS-International 2013 Top 30

1,719 New Researchers (75 countries). 9,089 Extracts, 87,292 Samples

Country Sample by Rank			Top Ranked Institution by Country	
Rank	**Sample**	**N**	**(Extracts)**	**N**
1	Brazil 2010	1,244	Harvard University – USA	296
2	Mexico 2010	1,198	Autonomous Univ. of Barcelona - Spain	239
3	Colombia 2005	878	Hong Kong Univ. of Science & Tech - China	221
4	Indonesia 2010	839	Univ. of the Witwatersrand - South Africa	215
5	Argentina 2010	806	Federal University Minas Gerais - Brazil	214
6	India 2004	798	University of Queensland - Australia	142
7	Chile 2002	743	Vienna Institute of Demography - Austria	142
8	South Africa 2007	647	National University of Malaysia	132
9	Canada 2001	644	Universidad de Montevideo - Uruguay	132
10	USA 2010	642	National University of Singapore	131
11	China 1990	633	Universidad Nacional de La Plata - Argentina	119
12	Viet Nam 2009	632	Pontificia Universidad Católica - Chile	106
13	Ecuador 2010	598	Paris School of Economics - France	105
14	Venezuela 2001	586	University College London - UK	104
15	Nicaragua 2005	584	Universität Tübingen - Germany	94
16	Uganda 2002	578	American University of Beirut - Lebanon	50
17	Bolivia 2001	577	Hebrew University – Israel	45
18	Peru 2007	576	World Health Organization - Switzerland	43
19	Uruguay 2006	569	National Research University - Russia	35
20	Spain 2001	562	Université de Montréal - Canada	34
21	El Salvador 2007	561	Université Catholique de Louvain - Belgium	29
22	Egypt 2006	552	University of Tokyo - Japan	29
23	Malawi 2008	551	University of Lodz - Poland	29
24	Costa Rica 2000	543	Università Degli Studi di Milano - Italy	28
25	Ghana 2000	525	Banco de la Republica - Colombia	24
26	Kenya 2009	509	UN-Habitat – Kenya	22
27	Panama 2010	509	Stockholm University - Sweden	22
28	Senegal 2002	497	University of Groningen - Netherlands	15
29	Philippines 2000	484	Chulalongkorn University - Thailand	15
30	Turkey 2000	478	Makerere University - Uganda	14

Source: IPUMS-International User Statistics Database, January 1, 2014. (Institution statistics exclude IPUMS's home, the University of Minnesota.)

Table 4. Usage of Select Variables in 50,181 Data Extracts Requested by Registered Users of IPUMS-International

Variables Requested	N
Extract requests	50,181
Samples requested	252,461
Total variables requested	1,796,196
Person-level variables	1,234,584
Detailed-level variables	264,580
Derived variables:	
Pointer ID	176,908
Technical variables:	
WTPER	49,540
WTHH	13,451
Household relationship pointers:	
SPLOC	11,189
MOMLOC	10,824
POPLOC	10,072
SPRULE	8,821
PARRULE	7,145
STEPMOM	6,974
STEPPOP	6,549
HEADLOC	4,013
Miscellaneous constructed variables:	
AGE2	13,768
SUBSAMP	8,090
BIRTHYR	5,637
MGRATEP	5,574
MGYRS1	5,135
DISABLE	4,365
NATION	4,196
COMPUTR	2,552
CELL (phone)	1,600
HOTWATR	1,596

Source: IPUMS-International User Statistics Database, January 1, 2014. (Institution statistics exclude IPUMs's home, the University of Minnesota.)

Tables 5: Thirty Most Commonly-Requested Variables by Record Type

	Person Record		Household Record	
Rank	Variable	N	Variable	N
1	EDATTAN	55,514	OWNRSHP	20,199
2	MARST	49,766	NCHILD	11,665
3	EMPSTAT	49,110	PERSONS	11,592
4	AGE	40,670	FAMSIZE	9,659
5	SEX	38,469	GQ	9,497
6	RELATE	37,822	NCHLT5	9,184
7	YRSCHL	19,393	ELDCH	8,798
8	SCHOOL	18,751	YNGCH	8,525
9	LIT	17,922	FAMUNIT	7,988
10	OCCISCO	17,710	WATSUP	7,510
11	URBAN	16,954	NFAMS	7,384
12	INDGEN	16,626	NCOUPLS	7,220
13	RELIG	16,128	ELECTRC	7,193
14	NATIVTY	14,991	ROOMS	7,005
15	CHBORN	14,844	TOILET	6,500
16	OCC	14,315	NMOTHRS	6,355
17	BPLCTRY	13,267	SEWAGE	6,133
18	IND	12,306	HHTYPE	5,912
19	CHSURV	9,922	NFATHRS	5,092
20	MGRATE5	9,326	PHONE	4,846
21	INCTOT	8,305	FUELCK	4,343
22	REGIONW	7,334	TV	4,036
23	CONSENS	7,254	WALL	3,928
24	CITIZEN	7,104	FLOOR	3,856
25	INCEARN	7,024	AUTOS	3,594
26	UNREL	6,243	KITCHEN	3,476
27	HRSWRK1	5,982	RADIO	3,457
28	POLY2ND	5,861	REFRIG	3,335
29	POLYMAL	5,835	BEDRMS	3,117
30	RACE	5,705	ROOF	3,008

Source: IPUMS-International User Statistics Database, January 1, 2014. (Institution statistics exclude the IPUMS's home, the University of Minnesota.)

Table 6. IPUMS Integrates Census Variables to Capture Common Concepts while Preserving Detail

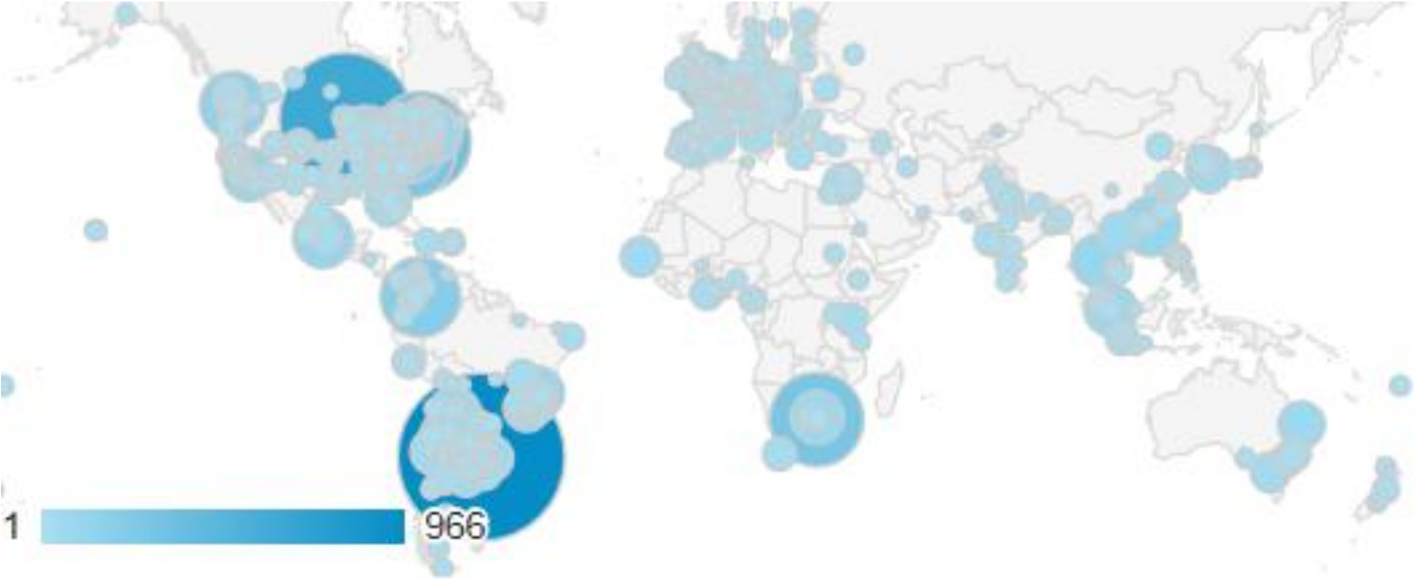
Educational attainment harmonized codes for the most recent sample of 16 large countries

("x" indicates that the code is present in the respective sample)

		Country (ISO 3166 code)	BR	CN	EG	FR	DE	IN	IR	MX	PK	PH	ZA	ES	SD	TH	US	VN
Code	**Variable Label**	**Sample year**	**00**	**90**	**06**	**06**	**87**	**04**	**06**	**06**	**98**	**00**	**07**	**01**	**08**	**00**	**05**	**09**
General (1 digit) Codes and Labels																		
0	NIU (not in universe)		x	x	x	x	x	·	x	x	x	x	x	x	x	x	x	x
1	Less than primary completed		x	x	x	x	·	x	x	x	x	x	x	x	x	x	x	x
2	Primary completed		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
3	Secondary completed		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
4	University completed		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
9	UNKNOWN/MISSING		·	·	x	·	x	x	x	x	x	x	x	·	x	x	·	·
Detailed (3 digit) Codes and Labels																		
0	NIU (not in universe)		x	x	x	x	x	·	x	x	x	x	x	x	x	x	x	x
100	LESS THAN PRIMARY COMPLETED		·	·	x	·	·	·	·	·	·	·	·	·	·	·	·	·
110	No schooling		x	x	·	x	·	x	x	x	x	x	x	x	x	x	x	x
120	Some primary		x	x	·	x	·	x	x	x	x	x	x	·	x	x	x	x
130	Primary (4 years)		x	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
PRIMARY COMPLETED, LESS THAN SECONDARY																		
	Primary completed																	
211	Primary (5 years)		·	·	·	·	·	x	x	·	·	·	·	x	·	·	·	x
212	Primary (6 years)		x	x	x	x	x	·	·	x	x	x	x	·	x	x	x	·
	Lower secondary completed																	
221	General and unspecified track		x	x	x	x	x	x	x	x	x	·	x	x	x	x	x	x
222	Technical track		·	·	·	x	·	·	·	x	·	·	·	·	·	·	·	·
SECONDARY COMPLETED																		
	General or unspecified track																	
311	General track completed		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
312	Some college/university		x	x	·	·	·	x	x	x	·	x	·	·	·	x	x	x
320	Technical track		·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
321	Secondary technical degree		·	x	·	·	x	·	x	x	·	·	·	x	·	x	·	x
322	Post-secondary technical education		·	x	x	·	x	x	·	x	x	x	·	x	x	x	·	·
400	UNIVERSITY COMPLETED		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
999	UNKNOWN/MISSING		·	·	x	·	x	x	x	x	x	x	x	·	x	x	·	·

Source: https://international.ipums.org/international-action/variables/EDATTAN#codes_section

Figure 2. Intense Traffic to IPUMS-International Website was Sparked by the Release of the Integrated Sample of the 2010 Census of Argentina



Buenos Aires led the world in traffic to the IPUMS-International website from August 1 to September 18, 2013, during the first 45 days following release of the 2010 census sample for Argentina. IPUMS-International acknowledges with thanks the Instituto Nacional de Estadística y Censos (INDEC) of Argentina for making the sample available without delay through IPUMS-International.
Analytical Data Source: Google Analytics, City Statistics: n=12,360