The Big Data Revolution: IPUMS–International. Trans–Border Access to Decades of Census Microdata Samples for Three–Fourths of the World and more

Robert McCaa[1]

Abstract

Over the past decade a revolution has occurred in the dissemination and analysis of census microdata. This paper discusses the IPUMS-International initiative to liberate census data for researchers world-wide without cost. Academic researchers and policy makers may access as many as 238 anonymized samples representing 74 countries and totaling over one-half billion person records. The database expands with the addition of 20-30 samples each year. Data are downloadable as extracts from the project website. To facilitate good use, both metadata and microdata are integrated. The analysis of 450 citations in the project bibliography reveals patterns in publications by country and topic.

Keywords: census microdata, integration, dissemination, metadata, IPUMS-International, samples.

IPUMS–Internacional y la revolución de los datos: Acceso a décadas de muestras de microdatos censales a escala mundial

Resumen

En la última década ha tenido lugar una revolución en el campo de la difusión y el análisis de microdatos censales. Este artículo presenta el proyecto IPUMS-Internacional que proporciona datos censales sin coste a investigadores de todo el mundo. Investigadores, profesores y expertos en políticas públicas tienen acceso a 238 muestras anonimizadas que representan 74 países y comprenden

1 Minnesota Population Center, University of Minnesota (rmccaa@umn.edu).

Fecha de recepción: abril de 2012. Fecha de aceptación: marzo de 2013
Revista de Demografía Histórica, XXX, I, 2012, segunda época, pp. 69–87

poco menos de 500 millones de registros individuales. La base de datos crece a razón de 20 a 30 muestras por año. Los datos se descargan directamente de la página web del proyecto. Para promover un buen uso, tanto los metadatos como los microdatos están integrados. El análisis de más 450 referencias bibliográficas muestra pautas de publicación por país y área temática.

Palabras clave: microdatos censales, integración, difusión, metadatos, IPUMS-Internacional, muestras.

IPUMS–International et la révolution des données: accès á des décades d'échantillons de microdonnées de recensements à l'échelle mondiale

Résumé

La dernière décennie a connu une révolution dans le domaine de la diffusion et de l'analyse de microdonnées de recensement. Cet article présente le projet IPUMS-Internacional dont l'objectif est de fournir des données de recensement sans coût aux chercheurs du monde entier. Les chercheurs, professeurs et experts en politiques publiques ont ainsi accès à 238 échantillons anonymisés qui représentent 74 pays et comprennent un peu moins de 500 millions de registres individuels. La base de données se développe au rythme de 20 à 30 échantillons par an. Ces données peuvent être téléchargées directement depuis la page Web du projet. Les métadonnées comme les microdonnées sont intégrées afin d'en faciliter l'usage. L'analyse de plus de 450 références bibliographiques montre les conventions de publication par pays et par domaine thématique.

Mots clés: microdonnées de recensement, intégration, diffusion, métadonnées, IPUMS-Internacional, échantillons.

INTRODUCTION[2]

The Big Data Revolution, foretold a decade ago in *Scandinavian Population Studies* (McCaa and Ruggles 2002), has arrived, but it is not

2 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.

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, 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 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, dotcoms beckon as a new job's 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. Computer tape was the medium of dissemination with sixty million integrated person records packed to a reel. Two years later, the Integrated Public Use Microdata Series (IPUMS) opened its first internet website, and dissemination by tape was quickly forgotten. Today, "IPUMS-USA" disseminates custom-tailored extracts of samples for any of the US 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 shortly after release by the Census Bureau. The updates include triennial and quinquennial versions of the ACS as well as annual.

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. 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.

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). The DwB initiative seeks to promote access and research for all kinds of European microdata. Worldwide, IPUMS-International seeks to engage with other national, regional and international academic, government and non-governmental partners.

In 2003, the North Atlantic Population Project, also hosted by the MPC, began with a focus on microdata from nineteenth century population censuses. At present NAPP disseminates integrated historical microdata for six countries—Canada, Great Britain, Iceland, Norway, Sweden, and the United States (see www.napp-data.org). Full-count microdata, including names of individuals, are available for eleven censuses, covering entire national populations. Person records total in the hundreds of millions. A NAPP ambition is to construct life histories for immigrants and non-immigrants alike through the full corpus of censuses over time and across borders for the entire North Atlantic region.

In 2012, the MPC launched the Terra Populus initiative to construct a global population and environment network. TerraPop seeks to harness census microdata to global-scale data on land use, land cover, climate change and more.

This paper focuses on the IPUMS-International "collaboratory" (Cogburn, 2003:89), specifically the functioning of trans-border dissemination of microdata and the exchange of research results. For orientation, readers are invited to examine the IPUMS websites. All pages are open-access, including dynamic metadata, census forms, field instructions, training manuals and the bibliographical database. To access microdata, interested researchers and students must register and agree to abide by strict terms of use. Readers whose official statistical authorities are not yet participating in the project are encouraged to lobby their agencies to cooperate with the IPUMS-International initiative.

1. IPUMS – INTERNATIONAL PARTNERS

Currently, 100 official statistical agencies participate in IPUMS-International, up from fewer than a dozen ten years ago. 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 (31): 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, 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 (20): Armenia, Bangladesh, Cambodia, China, India (Ministry of Statistics and Planning Implementation, not the Office of the Census Commission), Indonesia, Iran, Iraq, Israel, Jordan, Kyrgyzstan, Malaysia, Mongolia, Nepal, Pakistan, Palestine, Philippines, Thailand, Turkmenistan, Repiblic of Korea and Vietnam.
- Europe (20): Austria, Belarus, Bulgaria, Czech Republic, France, Germany, Greece, Hungary, Ireland, Italy, Netherlands, Poland, Portugal, Romania, Slovenia, Spain, Switzerland, Turkey, Ukraine, Russian Federation 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 eight largest with official statistical agencies yet to participate in the IPUMS-International initiative are: 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 these and others not yet inclined to cooperate.

Two-hundred and 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 integration and 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 for accredited researchers. Notable challenges remain for several historical censuses, specifically: Bangladesh (1981), Italy (1981 and 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

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

158 Countries/Territories						
Census Round	*Number Conducting a Census*	*% of World Population Enumerated*	*Country Microdata Inventory Extant*			
			2002	*2012*	*MPC*	*IPUMS*
1945-54	86	79	2	2	1	1
1955-64	116	87	27	24	21	14
1965-74	129	73	44	51	40	27
1975-84	138	96	54	75	55	34
1985-94	137	96	54	103	71	47
1995-2004	129	85	—	122	72	56
2005-14	149*	—	—	—	24	16

NOTE: 158 countries and territories with one million or more inhabitants in 2010.
For the 1960 round, two datasets thought to exist in 2002 are not yet usable--Canada, Philippines--, and for a third the only known copy was inadvertently destroyed (Austria).
* For the 2005-14 round the number of censuses is provisional (for dates, see: <http://unstats.un.org/unsd/demographic/sources/census/censusdates.htm>).
"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 with samples currently integrated and disseminated by IPUMS. Countries with more than one census per round are counted as one.
Sources: McCaa and Ruggles (2002), Table 1; and "IPUMS-International microdata inventory": <http://www.hist.umn.edu/~rmccaa/IPUMSI/census_microdata_inventory.htm>.

2. 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.[3] For

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

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[4] world-wide on identical terms.

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

4 Note that we refer to "bona fide researcher", not the bemusing term cited by Duncan et al (2011, p. 139).

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 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.

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).

3. 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
Top 20 Countries by Number of Extracts for Most Recent Sample – 2011

Rank	*Country*	*Sample %**	*Variables (n)**	*Sample years*	*Extracts (2011)*
1	Brazil	5	106	1960, 70, 80, 91, 2000	712
2	Mexico	10	120	1960p, 70, 90, 95, 2000, 05	626
3	United States	5	92	1960, 70, 80, 90, 2000, 05	554
4	Colombia	10	120	1964p, 72, 85, 93, 2005	516
5	South Africa	10	108	1996, 2001, 07	428
6	Argentina	10	84	1970, 80, 91, 2001	417
7	Canada	2.5	59	1971p, 81p, 91p, 2001p	409
8	Chile	10	93	1960, 70, 82, 92, 2002	393
9	France	33	94	1962, 68, 75, 82, 90, 99, 2006	380
10	Spain	5	99	1981, 91, 2001	366
11	India (surveys)	0.1	56	1983, 87, 93, 99, 2004	359
12	Venezuela	10	103	1971, 81, 90, 2001	354
13	Greece	10	89	1971, 81, 91, 2001	327
14	Kenya	5	81	1989, 99	326
15	Uganda	10	113	1991, 2002	324
16	China	1	63	1982, 90	321
17	Tanzania	10	90	1988, 2002	317
18	Austria	10	75	1971, 81, 91, 2001	310
19	Ghana	10	81	2000	309
20	Bolivia	10	91	1976, 92, 2001	301

NOTE: Extracts for the 2000 round or most recent sample for 55 countries disseminated Jan 1 – Dec 31, 2011.
*Refers to integrated variables for the sample, including IPUMS constructed variables.
"p" = person sample; otherwise samples are of households.

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 University/Research Institutions by Number of Extracts – 2011

Institution	*Extracts*
Columbia University	558
University of Hong Kong (Hong Kong, SAR)	309
University of Michigan	296
Arizona State University	256
Institute for Health Metrics & Evaluation (Seattle)	245
Universidad Nacional de La Plata (Argentina)	205
Harvard University	205
National University of Singapore (Singapore)	198
Centre of Demographic Studies (Barcelona, Spain)	185
Dartmouth College	177
Baruch College	152
University of Virginia	141
Vienna Institute of Demography (Austria)	138
United Nations-Habitat (Nairobi, Kenya)	134
University of Stirling (Scotland, UK)	110

Institution	*Extracts*
University of Pennsylvania	108
Institute of Political Studies (Paris, France)	107
Vanderbilt University	105
University of Colorado at Boulder	90
University of Cape Town (South Africa)	88
University of Chicago	86
University of Guelph (Canada)	85
World Bank	81
University of California at Berkeley	77
Brigham Young University	76
Pompeu Fabra University (Barcelona, Spain)	75
Pew Research Center	74
Inter-American Development Bank	73
University of Groningen (Netherlands)	72
Kenyon College	70
London School of Economics (UK)	62
Bocconi University (Milan, Italy)	59
Indiana University	57

SOURCE: IPUMS-International User Statistics Database, January 1, 2012 (list excludes IPUMS's home, the University of Minnesota)

4. 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 harmonized codes: 16 large countries, most recent sample in IPUMS-International "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	*Label* *Sample year*	*00*	*90*	*06*	*06*	*87*	*04*	*06*	*00*	*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																	
000	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/codes.do?mnemonic=EDATTAN>.

5. 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 according to 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.

6. 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.

Participating statistical agencies are invited to entrust metadata and microdata for the 2010 census round at their earliest convenience. Agencies not yet cooperating with the IPUMS-International initiative are invited to consider doing so. Researchers who have yet to access the IPUMS microdata are invited to peruse the open-access metadata and submit an application should their research needs require. Meanwhile the Big Data Revolution continues.

BIBLIOGRAFÍA

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 DOMINGO-FERRER, J. and TINNIRELLO, I. (eds.) (2012): *Privacy in Statistical Data 2012, LNCS*, vol 7556.

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

ESTEVE, A. and SOBEK, M. (2003): "Challenges and methods of international census harmonization", *Historical Methods* 36: p.p. 66-79.

LOHR, S. (2012): "New U.S. Research Will Aim at Flood of Digital Data", *New York Times*, March 29, p. B2.

McCAA, R. and ESTEVE, A. (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, R., RUGGLES, S., DAVERN, M., SWENSON, T. and PALIPUDI, K. M. (2006): "IPUMS-International High Precision Popoulation Census Microdata Samples: Balancing the Privacy-Quality Tradeoff by Means of Restricted Access Extracts." in *Privacy in Statistical Databases*. New York: Springer, pp. 375-382.

McCAA, R. and RUGGLES, S. (2002): "The census in global perspective and the coming microdata revolution", *Scandinavian Population Studies* 13: 7-30.

MEIER, A., McCAA, R. and LAM, D. (2011): "Creating statistically literate global citizens: The use of IPUMS-International integrated census microdata in teaching", *Statistical Journal of the IAOS* 27, p.p. 145-156.

SOBEK, M., CLEVELAND, L., FLOOD, S., KING, M., RUGGLES, S. and SCHROEDER, M. (2011): "Big Data: Large-Scale Historical Infrastructure from the Minnesota Population Center", *Historical Methods* 44: 61-68.

THOROGOOD, D. (1999): "Statistical Confidentiality at the European Level", *Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Thessaloniki, March.

VAN HOOK, J. and GLICK, J. (2007): "Immigration and Living Arrangements: Moving Beyond Economic Need versus Acculturation", *Demography* 44, p.p. 225-249.