IPUMS-International: A Decade of Progress and Challenges and Their Relevance for Disseminating Integrated, High Precision Samples of the Population Censuses of India.

by

Robert McCaa, Steven Ruggles, Matthew Sobek, and Lara Cleveland
University of Minnesota Population Center. Minneapolis, MN . U.S.A

Abstract.

IPUMS-International is a global initiative begun in 1999 to preserve, integrate and disseminate sample extracts of the world's census microdata. 97 official statistical agencies—including for India the Ministry of Statistics & Programme Implementation, but not the Office of the Registrar General—have endorsed the IPUMS-I protocols, encompassing 87% of the globe's population. Researchers world-wide may download free of cost from www.ipums.org/international custom-tailored extracts of samples for more than two hundred censuses covering as many as five decades. Led by the University of Minnesota Population Center (MPC), the project is funded by the National Science Foundation and the National Institutes of Health (USA).

Challenges faced by the project may be grouped into four types: financial, legal, technical and administrative. Specifically, our list of the most notable obstacles include: obtaining official consent, recovering data from old tapes, documenting the microdata, drawing, editing and confidentializing high-precision household samples, integrating the metadata and microdata, managing access to the microdata, promoting quality research, sharing fruits of the research, etc.

Despite these challenges and more, each year the database grows with the addition of samples for 15-25 censuses, representing a half dozen or so countries. Currently, extracts of 212 samples, encompassing 69 countries and totaling over 480 million unit (person) records, are being disseminated to more than six thousand registered researchers world-wide. These figures are expected to double over the current decade. Meanwhile, a great outpouring of research has produced hundreds of publications, including a couple of dozen books and doctoral dissertations based in-whole or in-part on IPUMS-International integrated census samples.

The censuses of India are not presently represented in the database--despite their fame and their distinguished history. The purpose of this paper is to stress the importance of integrating high precision samples of the 1991, 2001, and 2011 censuses of India into the IPUMS-International database. The paper concludes with a comparison of definitions and concepts in population censuses of India and 16 Asian countries already represented in the database. Our analysis of the 1991-2011 Indian censuses shows that they are second to none in terms of comprehensiveness as well as in detail and consistency of questions posed—all the more remarkable given that the censuses of India rank among the world's largest peace-time, democratic undertakings.

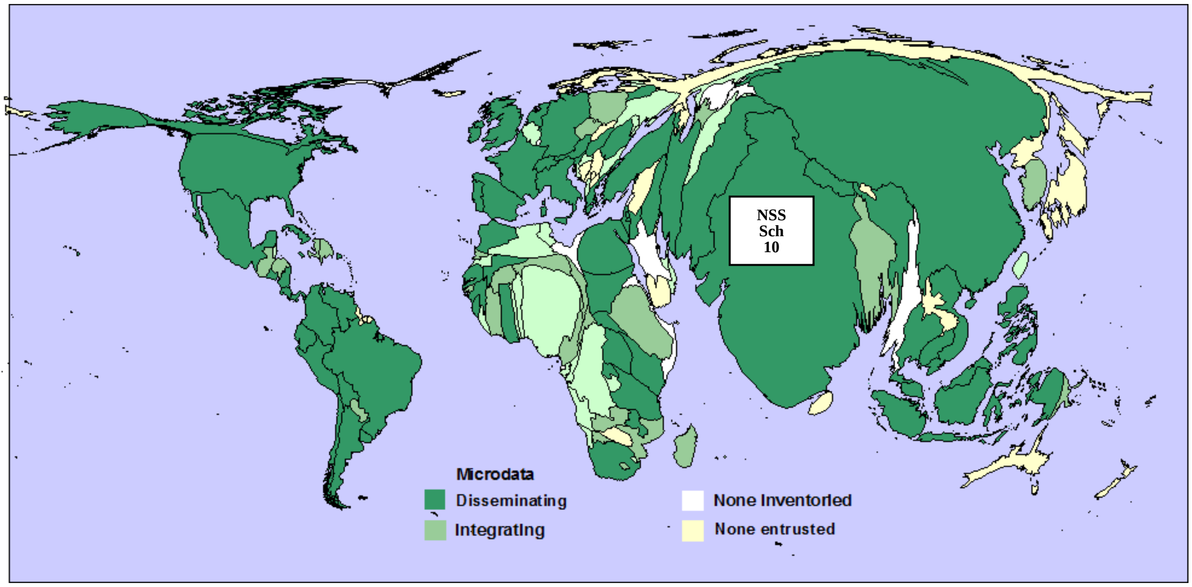
Paper presented at the Census 2011 Conference organized by Centre for the Study of Regional Development, Jawaharlal Nehru University September 6-7, 2012

A decade of progress.

IPUMS-International is a global collaboration of universities, national statistical authorities, data repositories, and research centers to archive, integrate, and disseminate census microdata. Currently 97 statistical offices are participating in the project, encompassing 87% of the globe's population (Figure 1). In this figure the three shades of green--disseminating (dark), integrating (medium), and negotiating (light): India is represented in the database, not by census samples as in the case of all other countries, but by National Sample Surveys of Employment and Unemployment (NSS Schedule 10).

Figure 1

The IPUMS-International collaboratory, August 2012
Country cartogram weighted by population size



Founded in 1999 and led by the University of Minnesota Population Center (MPC), the project currently disseminates 212 confidentialized, integrated population samples, representing 69 countries and totaling almost one-half billion person records (Table 1).

Each year the database is updated with samples from the latest 2010 round censuses and for five to ten additional countries as integration of microdata is completed. In 2013, we expect to add samples for as many as ten countries. Candidates include: Bangladesh (1981-2011), Brazil (2010), Cameroun (1976-2005), Ecuador (2010), Fiji Islands (1966-2007), Ghana (1984, 2010), Israel (1961, 2008), Kenya (1969, 1979, 2009), Kyrgyz Republic (2009), Panama (2010), and South Africa (2011). By the end of our second decade, census samples for more than 100 countries are likely to become available through the IPUMS-International and partner portals.

Integrated microdata are available free-of-cost to researchers world-wide on a restricted access basis. To gain access, the applicant must establish research bona fides, demonstrate both a need and an ability to use the microdata, and pledge to respect stringent terms of use. The registration form is lengthy and intimidating, precisely to dissuade unqualified applicants. Qualified applicants, once approved, are authorized to submit requests via the internet by means of a series of point-and-click menus selecting specific country(ies), census(es), subpopulation(s), and variables. Each request automatically generates an "extract", an integrated, custom-tailored file of pooled microdata—regardless of the number of censuses or countries selected. The researcher then downloads the extract for analysis using whatever hardware and software is preferred.

Table 1: **Integrated population microdata samples currently disseminated (the chronological scope of the series for each country is indicated in parentheses)**

Africa 28 samples	*Egypt (1996-2006), Ghana (2000, 2010 in preparation), Guinea (1983-1996), Kenya (1989-1999 plus 1969, 1979, 2009 in preparation), *Malawi (1987-2008), Mali (1987-1998), Morocco (1982-2004), Rwanda (1991-2002), Senegal (1988-2002), Sierra Leone (2004), *South Africa (1996-2007, 2011 in preparation), *South Sudan (2008), *Sudan (2008), Tanzania (1988-2002), Uganda (1991-2002)
Americas 81samples	Argentina (1970-2001), Bolivia (1976-2001), Brazil (1960-2000 plus 2010 in preparation), Canada (1971-2001), Chile (1960-2002), *Colombia (1964-2005), Costa Rica (1963-2000), Cuba (2002), Ecuador (1962-2001 plus 2010 in preparation), *El Salvador (1992-2007), Jamaica (1982-2001), *Mexico (1960-2010), *Nicaragua (1971-2005), Panama (1960-2000), *Peru (1993-2007), *Puerto Rico (1970-2005), Saint Lucia (1980-1991), *USA (1960-2005 plus 2010 in preparation), *Uruguay (1962-2006), Venezuela (1961-2001)
Asia and Oceania 47 samples	*Cambodia (1998-2008), China (1982-1990), India (1983-2005 NSS Schedule 10 surveys), *Indonesia (1971-2010), *Iran (2006), Iraq (1997), Israel (1972-1995, 2008 in preparation), Jordan (2004), Kyrgyz Republic (1999, 2009 in preparation), Malaysia (1971-2001), Mongolia (1989-2000), Nepal (2001), Pakistan (1973-1998), *Palestine (1997-2007), Philippines (1990-2000), Thailand (1970-2000), *Vietnam (1989-2009).
Europe 56 samples	Armenia (2001), Austria (1971-2001), Belarus (1999), *France (1962-2008), Germany (1970-1987—includes GDR and FRG—plus 1991-2011 in preparation), Greece (1971-2001), Hungary (1970-2001), *Ireland (1971-2006), Italy (2001 plus 1981-1991 in preparation), the Netherlands (1960-2001), Portugal (1981-2001), Romania (1977-2002), Slovenia (2001), Spain (1981-2001), Switzerland (1970-2000), Turkey (1990-2000), the United Kingdom (1991-2001 plus 1961-1981 in preparation)

Note: * = sample for 2010 round population census already integrated into IPUMS-International

Samples for the 2010 round of censuses (2005-2014) are assigned the highest priority for integration. As soon as data processing is completed by the national census agency and entrusted to the Minnesota Population Center, we begin integrating the documentation. Once that step is completed, we proceed to integrate the microdata. 2010 round samples for eighteen countries are already incorporated into the database (denoted by an asterisk in Table 1). To facilitate the analysis of change over time, our goal is to integrate a complete chronological series of census microdata for each country, beginning with the earliest census for which microdata exist or can be reconstructed and continuing through the most recent census. Of the 69 countries currently in the database, 39 are represented by samples stretching over three decades or more.

India is currently represented in the database by series of five employment and unemployment household surveys (Schedule 10 of the National Sample Survey Organization), dating from 1983 through 2005. A sixth is planned for launch in 2014. The strengths of NSS surveys are well known: extent of topics, consistency in questions, continuity over several decades, quality of field operations and excellence of data processing facilities. However, Schedule 10 samples suffer from two serious weaknesses, in comparison to census samples: lack of detailed coverage at low levels of geography and relatively low precision (0.5%)—a paltry 600,000 records to represent a billion people. In contrast, for Indonesia, more than 23 million person records make up the ten percent sample of the 2010 census, the most recent sample in the series of nine (5 censuses and 4 surveys) integrated into the IPUMS-International database.

Only population censuses offer the necessary geographic detail and dense coverage. Hopefully the Office of the Registrar General of India will agree to a plan to entrust to the Minnesota Population Center high-precision household samples for each of the censuses of India for which microdata are extant. Indian censuses are famous both for their lengthy, distinguished history as well as for their contributions to the fields of population studies, human geography and the social and statistical sciences in general. This paper will demonstrate that the content of Indian population censuses for the years 1991-2011 is second to none in terms of the rigor and consistency of concepts and definitions. Before discussing the relevance of the IPUMS-International initiative to Indian census microdata, first we summarize some of the major challenges faced by the project and our solutions for overcoming them.

Challenges

We faced four types of challenges—financial, legal, technical and administrative. Fortunately all proved to be more easily solved than might be imagined given the sensitive nature of census microdata and the global vision of the IPUMS-International undertaking.

Financial challenges were the least troublesome. In 1998, the National Science Foundation Social and Behavioral Branch issued a call for proposals to construct global social science infrastructure. Our proposal for a decade-long project for twenty-one countries was the only project graced with funding. Nonetheless, there was great skepticism regarding the prospects for success. Our budget was slashed by two-thirds, and the project reduced to a mere pilot of eight countries: Brazil, China, Colombia, France, Kenya, Mexico, USA, and Vietnam. Three years later, with the first signs of the pilot's success, the National Institutes of Health (NIH) funded a five-year long sister project focused on Latin America. In turn, a third project, for Europe was funded in 2004. Renewal grants have continued in an uninterrupted series from both NSF and NIH—thanks to highly favorable assessments by peer-review boards. Recently, the geographical scope of IPUMS-Europe was extended to Eur-Asia.

Legal challenges constituted the biggest obstacle. Census microdata are highly sensitive because of the need to protect privacy and statistical confidentiality. In addition in some countries, census data are considered state secrets. Then too, some census agencies hide behind the law or the absence of law, fearful that disseminating microdata may reveal errors, omissions, inconsistencies, fraud or other embarrassing details.

In 1994 the United Nations Statistical Commission, in a big step forward, recognized anonymized microdata samples as statistical products that could be disseminated "to honor citizen's entitlement to public information". By 2001, the International Monetary Fund's

General Data Dissemination System revealed that three-fourths of member states disseminated census microdata in one guise or another.

We studied the best practices and sought to incorporate them into two legal documents: one to safeguard the interests of the owners of the microdata, the official census agencies, and the other to protect the usage of the microdata. The former is governed by a uniform Memorandum of Understanding (MOU) consisting of ten protocols (later expanded to eleven—see Appendix A) and the latter by a lengthy terms of use license (see Appendix B). Attorneys of the University of Minnesota meticulously reviewed both documents for approval and continue to exercise close legal supervision of transactions with each National Statistical Office partner as well as the user license.

When we began, few statistical agencies welcomed the intrusion of academics—much less historians!—into official circles. Over time, a growing number of agencies entrusted their most treasured microdata to the project. A major turning point came in 2007 with a week-long, on-site inspection of our operations by one of the most respected experts in the field of microdata management, the former Australian Statistician Dennis Trewin. Mr. Trewin concluded his report as follows (Trewin 2007):

> Without question IPUMS-International meets the four Core Principles outlined in CES [Conference of European Statisticians] (2007). It is cited in CES (2007) as a Case Study of good practice. This review confirms its status as good practice for Data Repositories. Indeed it is likely to provide the best practice for a Data Repository for international statistical data [emphasis added].

Since 2007, the number of statistical agencies endorsing the project protocols has more than doubled. Forty of the fifty most populous countries participate in IPUMS-International. The ten largest whose census agencies are yet to embrace the project protocols are: India, Nigeria, the Russian Federation, Japan, Algeria, Korea (RO), Saudi Arabia, Korea (PDR), Yemen, and Syria. Negotiations are on-going with these and smaller countries, as conditions permit.

Technical challenges.

When we began, integrating census microdata and metadata on a global scale seemed impossible. Indeed, it was for this reason that the NSF scaled back the first grant to a pilot. Fortunately in the late 1990s the internet was maturing rapidly and new tools were emerging for managing access not only to massive amounts of quantitative data but more importantly to manage enormous amounts of text. Our vision was to design a scalable system that could encompass all the world's census microdata and documentation. We faced four types of major technical challenges: microdata recovery, metadata (documentation) recovery, integration and dissemination.

Microdata recovery. The first task was to recover historical microdata from tapes, cartridges, floppy discs and other out-dated media. Early on we located a commercial data recovery company, Muller Media Conversions (MMC), that has old equipment, software and technical expertise to rejuvenate media that are seemingly impossible to decipher. As an example, consider the recovery of the 1981 census of Bangladesh. This remains our costliest, but most successful project to date. The data were stored on 279 9-track tapes of different manufactures. Many of the tapes had to be “baked” to force the magnetic medium to re-attach to the plastic backing. Great care was taken to avoid melting the tape. Others were coated in mould. We succeeded in recovering a one percent sample for almost the entire country, plus full count microdata for approximately half. The recovery effort is being continued by staff of the Bangladesh Bureau of Statistics with technical assistance provided by MMC.

In other countries, the microdata were secure, but substantial effort was required to develop documentation. This is the case of the United Kingdom, where the microdata for censuses of 1961, 1971, and 1981 are secure, but unusable. Last month, a project finally got underway led by the UK Data Archive to reconstruct the databases for each of the censuses and prepare high precision samples for both UK and IPUMS users. The expected completion date is 2014. In most countries, the task of recovering the microdata is left to the IPUMS project team.

Metadata (documentation) recovery and archiving. Metadata preservation progresses in two major areas. First is the collection and processing of metadata related to the census microdata entrusted to the Minnesota Population Center. These metadata include data dictionaries, codebooks, enumeration forms, enumerator instructions, and other technical materials. These documents are scanned, translated into English where necessary, and their contents are reformatted into a uniform structure. These metadata are used to provide the source material for harmonization as well as to populate the metadata specification that drives the IPUMS-International dissemination system.

The second area of metadata archiving activity is the cataloging and preservation of over 25,000 documents related to international censuses covering the period 1960 to the present. These documents provide a rich source of material that cover technical specifications, preparations for census activities, involvement of international agencies, internal organization, promotional materials, and subsequent publications and reports. The collection consists of over 10,000 items from the United Nations Statistical Division, 8,000 documents from the United States Census Bureau International Collection, plus contributions of materials from regional repositories, such as Centro Latinoamericano y Caribeño de Demografía (CELADE), Centre Population and Développement (CEPED), as well as from numerous national statistical organizations and private collections (e.g., Rand-McNally Inc.). These collections provide broad access to the core metadata required to understand national census samples. Many of the source documents are in poor condition due to age, original publication quality, or lack of preservation. Scanning preserves the content and provides a means of access and dissemination for both the IPUMS project and the National Statistical Offices of the originating countries.

Integration. IPUMS-International is *not* simply a conduit for passing census samples from National Statistical Offices along to researchers. Instead, two or more years of labor are typically expended by a large technical staff working at the MPC to integrate census microdata and metadata for dissemination. In the twenty-first century, handing along a copy of the source microdata and a data dictionary is *not* sufficient for high quality research. To meet this need, the MPC has developed a routine of five steps for processing each census before microdata are disseminated by IPUMS-International.

1. Confirm the integrity and validity of the source microdata and metadata
2. Construct a high precision, geographically stratified, confidentialized sample on which all subsequent work is based
3. Integrate the microdata, variable-by-variable, code-by-code, using, where possible, international standards (ISCO, ISIC, ISCED, etc.) and, where necessary, developing new composite coding schemes (e.g., relationship to head, marital status, etc.).
4. Integrate the metadata. Write succinct descriptions of integrated variables and codes as well as comparability discussions to high-light significant incompatibilities.
5. Rigorously test the integrity and validity of the integrated microdata sample and metadata

For the sake of brevity, this paper discusses only the core innovations, items 2-4.

Constructing a high precision, geographically stratified, confidentialized sample for each census. Of the 212 samples currently in the database, 127 (60%) were constructed to IPUMS specifications:

1. High precision, geographical stratification with the household as the sample unit.
2. Systematic sample design, after a random start:
   for private households, every nth household;
   for group dwellings (hotels, hospitals, boarding schools, etc.), every nth person.
3. 10% sample density

Sixty-eight samples consist solely of the extant microdata, where no other design can be used—either because the complete dataset no longer survives or because the sample was taken “in the field,” often using a longer questionnaire. Of the remaining 17, where a sample could be drawn to IPUMS specifications, five are being re-drawn and others are under consideration.

Once the sample is constructed, it must be confidentialized before integration can proceed. Microdata entrusted to the IPUMS-International project are subjected to strong technical statistical disclosure controls providing greater protections for the group of statistical offices as a whole than for any single agency that chooses to “go it alone” (McCaa, Ruggles and Sobek 2010). Moreover, confidentializing can easily go awry, even for statistical agencies with decades of experience in confidentializing microdata (Alexander, Davern, and Stevenson 2010).

There are four steps to the process of confidentializing samples in the IPUMS-International database. First, the most important technical statistical disclosure control is the suppression of records by the construction of the sample. All the values in the records outside the sample are suppressed. Second is the suppression of names and low level geographical detail. Each statistical authority balances the confidentiality/utility trade-off by instructing the project as to the minimum threshold for identifiable geographical units. For many countries, the threshold is commonly set at 20,000 inhabitants. Others place it as high as 100,000 (United States) or in the most extreme case (Netherlands) all administrative geography is suppressed. Third, in consultation with the national statistical office, some variables may be top-coded, others may be subjected to global recoding, the deletion of digits for hierarchical variables (occupation, industry, geography), or the suppression of a variable entirely. Fourth, additional statistical disclosure protections are provided by randomly ordering the records and swapping the geographical identifiers of an undisclosed number of households. This means that no one can allege with certainty that an individual or household has been identified.

Integrating the microdata. The principal benefit of IPUMS to researchers and NSOs alike is the integration of a complete, chronological series of microdata samples for each country—typically beginning with the earliest census for which microdata exist or are recoverable and continuing through the 2010 census round. Samples are integrated both chronologically and cross-nationally.

The basic goal of the IPUMS-I harmonization effort is to simplify the use of the microdata and metadata while losing no meaningful information. This is challenging because to make the microdata simple for comparative analysis across time and space, it is necessary to develop comparable coding schemes. Microdata are integrated so that identical concepts have identical codes. To avoid the loss of important information for those samples that have even more detail, IPUMS-I uses a composite coding strategy 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, which we call 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 samples. Trailing digits provide detail that is only rarely available. No information for a digit is denoted by zero.

As an example of the IPUMS method of integrating variables, consider the concept "educational attainment," with 37,725 extracts (June 2012), 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 a missing data code (9) and, for children too young to attend school or others to whom the question was not addressed, a zero ("not in universe"), is assigned. Some 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 understand these and thousands of other details.

Appendix C illustrates the detailed and general coding schemes for the educational attainment variable for 15 countries, represented by the most recent census integrated into the IPUMS-I database. A similar table for any combination of countries or censuses is generated from the IPUMS-I home page by first clicking "Browse and Select Data". At this point, to avoid displaying metadata for all samples in the database, click "Select Samples," one or more countries or censuses as desired, and click "submit sample selections". To select the education attainment variable, mouse-over "Person," click "Education," and then on the "EDATTAN" line click "Codes". The display will indicate the general, one digit codes for the "Educational Attainment, international recode" for all countries and census years selected. "X" indicates that the corresponding code is available for the country and census indicated. Clicking "detailed codes" generates the detailed, three digit codes, as in Appendix C. Click "Case-count view" to change the “X” to simple, un-weighted frequencies for each code in the selected samples, as seen in Appendix C. Metadata of this type is prepared for every integrated variable in the database. All IPUMS-I metadata are open to everyone. No registration is required.

Integrating the metadata. To facilitate research we integrate into a single database both the microdata and the metadata (documentation) of all censuses for all countries. The metadata includes detailed descriptions of each census, sample and variable. IPUMS-International metadata offer succinct descriptions of each census in the database, listing the title, year, universe, de jure/de facto, enumeration unit, official census day, forms, field work period and type, respondent, and estimates of undercount, where available. Images of census enumeration forms and instructions manuals are available in the official language and the text in English, translated when necessary. Each sample is described with regard to source, sample design, sampling unit, sample fraction, number of unit records, sample weights, dwelling or housing units, vacant dwellings, households, group quarters and special populations.

IPUMS metadata define each integrated variable and describe basic characteristics: availability by census year and country, universe of the variable or question, codes, questionnaire text, and non-harmonized variables on which the microdata integration is based. Comparability discussions summarize the most important similarities and differences in definitions for each variable, including country or census specific information noting departures from standard practice. The purpose of these discussions is to highlight important contrasts. Clicking “Questionnaire text” on the variables page leads to source questions and corresponding instructions in English for each selected census. Additional clicks yield views of the original documentation in image form so that researchers may study lay-out and actual wording in the official language.

Researchers navigate the integrated metadata to readily examine census questions and instruction manuals for any combination of countries and census years. The MPC integration team applies XML tags to the census documents, associating variables in the microdata with concepts in the source documents. The tagged material is then imported into the IPUMS database. Once this step is completed, metadata may be retrieved dynamically for any combination of countries and census years, variable-by-variable. Initially this tool was developed to speed the work of the integration team at the MPC. Once the analytical power of the tool became apparent, we harnessed it to the web-site, to facilitate dynamic access to the metadata by the public.

For each microdata extract request, the IPUMS system automatically generates a comprehensive archival quality DDI (Document Data Initiative) compatible metadata dictionary, custom-tailored to the precise specifications of the researcher.

Administrative Challenges.

The project faces three kinds of management challenges: access, dissemination, and research results. We designed three web-pages to manage these problems:

1. Apply for access: https://international.ipums.org/international-action/register/new
2. Select data: https://international.ipums.org/international-action/variables/group
3. Bibliography: http://bibliography.ipums.org/

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

Agreeing to the conditions of use binds both the researcher and the researcher’s institution. The Legal Counsel of the University of Minnesota is poised to strike at the first indication of misuse. Both the individual researcher and the researcher’s institution are responsible for maintaining security and enforcing the license agreement. Violations are likely to lead to sanctions against both the individual and the offending researcher’s institution. A violation by a single user may suspend access to all users at that institution. Researchers at an embargoed institution may be sanctioned to undergo remedial training for the protection of human subjects so that the institution may regain its accreditation for handling sensitive microdata.

IPUMS-I procedures resolve the conundrum of managing the broadest possible access to sensitive microdata while protecting statistical confidentiality. Many statistical agencies have long wanted to make census microdata available to researchers, but lack the substantial material and human resources required to implement and manage secure systems.

The fact that IPUMS-International distributes microdata electronically as custom extracts, tailored as to country(ies), census year(s), subpopulation(s), and variables, according to the individual needs of the researcher, provides additional incentives for users to jealously guard the microdata. Since complete datasets are not distributed on CDs or any other media, the temptation to share microdata with unauthorized individuals is greatly reduced.

Google Analytics suggests that the IPUMS registration form alone is a substantial deterrent to casual users. Over a recent twelve month period, 5,593 views of the registration page yielded only 1,057 completed applications. One reason for the large drop-off is that the registration form is a daunting deterrent to the statistically naive.

A qualified researcher, regardless of the time required to fill-out the form, will readily agree to these conditions and meticulously provide the requested information, while the unqualified—faced with identifying by name the Human Subjects Protection Committee of his or her institution, supervisor, and website listing the individual’s institutional affiliation as well as describing the research project for which the microdata are to be used—will not complete the form at all. Incomplete forms are automatically rejected by the web page controls. It is impossible to submit an incomplete application. The daunting detail required to complete the form leads to self denial by the casual visitor.

Of the 1,057 completed applications noted above, a mere 46 were denied. A decade ago, when the form was simple, the denial rate averaged one-third. By re-designing the form, we substantially reduced the work-load of the vetting process. Once the completed registration is submitted, applicants are carefully vetted to prevent access to researchers who are unqualified or who lack a research need. For a majority of denials the currently disseminated census microdata were not suitable for the proposed research.

Despite the stringent conditions of use and restrictions more than six thousand researchers—representing over 100 countries and 900 institutions—are approved for access to the IPUMS-I database. One-third of IPUMS-I users request access to microdata for a single country, usually their own. A large fraction are studying abroad and are seeking access to data for their country of birth.

Managing dissemination of microdata. IPUMS disseminates extracts, custom-tailored to the precise research needs of each user. The average IPUMS extract last year consisted of a mere 35 variables, including 6 technical variables that are automatically included with each extract.

This contrasts with the practices of most statistical offices where microdata for a census are disseminated as single set, consisting of a data dictionary and an entire sample containing all variables and all person records. Typically, under this old-fashioned approach, when requests are fulfilled, each researcher receives exactly the same set of data and documentation. The temptation for passing these datasets on to unauthorized users is obvious.

With IPUMS-I no two extracts are alike. The fact that IPUMS-International distributes microdata electronically as custom extracts, tailored as to country(ies), census year(s), subpopulation(s), and variables, according to the individual needs of the researcher, provides additional incentives for jealously guarding extracts. Since complete datasets are not distributed on CD or other media, the inclination of researchers to share data with unauthorized individuals is greatly reduced.

Given the massive size of the IPUMS-International database, disseminating the full set of variables and unvarying size of samples is impractical. Selections are made by a series of point-and-click screens. To facilitate the selection process, metadata are readily available to surf the documentation in any sequence desired without interrupting the extraction process. To construct an extract the researcher selects:

- 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 old. Note that there is an option for selecting both the individual as well as all co-resident persons enumerated 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 single, pooled dataset containing the microdata, the corresponding set of DDI compatible metadata as well as user-selected SPSS, SAS or STATA system files and codebooks. Copies of original source metadata are available from the web-site, as well as integrated metadata in interactive form. These also may be downloaded freely.

The IPUMS extract engine adds even more value to each extract by means of three unique tools:

a. **Select cases.** This feature filters samples to select precisely the cases of research interest. For example, if the one wishes to research only economically active females, aged 15-19, born in a country different from the current country of residence, the IPUMS extract engine will generate a dataset precisely to these characteristics. Moreover, with a single click by the user, the extract engine will include all persons in the household with the selected individual.

b. **Attach characteristics.** By simply clicking the attach characteristics "change" option, the extract engine attaches variables of mothers, fathers, spouses and household heads to co-resident individuals. This feature facilitates the analysis of children by characteristics of their mothers, fathers, and/or household heads (Sobek and Kennedy 2009). The feature is useful for analyzing own-child fertility, marital homogamy, and a host of other topics where the joint-characteristics of two or more members of the household are required. For example, for the recent international seminar on "Cross Border Marriages" (Seoul, Republic of Korea, October 2011), it was easy to generate a dataset from the IPUMS-I website for 51 countries representing 12 million foreign born individuals married or in union with co-resident native born spouses (Esteve, Garcia and McCaa, 2011; cited in The Economist [Parker], 2011).

c. **Customize sample size.** The extract system offers a tool to customize the size of each sample in terms of the number or percent of persons or households. If the researcher desires a sample of only 50,000 households, simply enter "50" in the corresponding table and the extract engine will construct a systematic sample of the desired size. The appropriate weights (expansion factors) are automatically computed on-the-fly and included in the extract.

In 2005, at the UN-ECE Expert Group Meeting on Statistical Data Confidentiality in Geneva we explained the IPUMS-International data dissemination security procedures 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 file, decompress it and proceed with the analysis using the supplied integrated metadata consisting of variable names and labels.

This method of dissemination continues to weather the test of time, and indeed as usage soars, the rapid acceleration of internet transmission speeds has validated IPUMS-I security protocols. A record is kept of each extract so that if there is need, the extract may be replicated and provided either to the researcher or indeed any researcher who may wish to challenge or replicate findings. The National Science Foundation was particularly insistent on this feature as a means of reducing research fraud and the temptation to exaggerate findings.

We also share this technology with our partners to construct customized regional portals. The Autonomous University of Barcelona manages a portal for access to integrated samples for European countries and the African Cenre for Statistics hosts a portal for access to African census microdata.

Managing research results (bibliography). A great outpouring of research based in-whole or in-part on integrated samples disseminated by IPUMS-International has produced hundreds of publications, including a couple of dozen books and doctoral dissertations. Researchers cite their research results in the IPUMS open-access bibliography: http://bibliography.ipums.org/. Links are provided to abstracts and to the publications themselves, where copies are available on the web. The bibliography currently lists over 600 entries under the IPUMS-International tag of which a mere thirteen cite the NSS samples of India. Nine are comparative, analyzing at least two or as many as 47 countries. Topics include studies of marriage patterns, assortative mating, equitable development, sex selection, structural change, urban and regional population growth, over-education in developing economies, inequality in amenities, age inequality of spouses, gender inequality, global migration, etc.

Relevance to Population Censuses of India.

The National Sample Surveys of India rank a respectable eleventh in usage by IPUMS-International researchers for 2011. Given the specialized content of the Schedule 10 samples and their small size their high ranking is remarkable. Nonetheless, without a doubt the surveys are no substitute for census microdata. Of course the surveys are not intended to be censuses. They are integrated into the IPUMS database because no census samples are available for India, and the Ministry of Statistics and Planning Implementation generously consented to the integration of the Employment samples.

A significant shortcoming of the Surveys is content. Of the 35 most common questions in censuses of ten Asian countries in the IPUMS-International database, the NSS Schedule 10 Employment series contain a mere sixteen (see Appendix D). Lacking are questions on nativity, fertility, country of birth, migration, and disability. The 1991-2011 population censuses of India contain all these questions and more, ranking at the very top in terms of the comprehensiveness and consistency.[1] Twenty-six of the 35 questions listed in Appendix D are present in the 1991 census of India. The 2001 enumeration surpassed this mark by the addition of a module on disability status. The 2011 census added questions on birth year and month. Moreover, all three censuses include such important additional topics as scheduled caste/tribe, language, particulars of work, and place of birth and last residence. Many of these are absent from census questionnaires of other Asian countries.

A second weakness of the Employment Surveys is the absence of geographical detail. This defect could also be remedied by high precision census samples. The Surveys are limited to some 600,000 unit records. With a sample density of less than 0.1%, they do not pretend to offer statistics below the state level. High-precision population census samples, on the other hand, make it possible to drill-down to districts and modest sized towns of 100, 50 and even 20 thousand inhabitants.

[1] The 1981 census of India may not be relevant for this discussion because existence of the microdata is doubtful. In the late 1980s a five percent sample was constructed for research purposes, but as far as can be determined no researcher succeeded in making use of or preserving the microdata (Premi 2001). Today, if a copy of the 1981 sample should be recovered, it would constitute a treasure, not only for India, but for the world.

Ten percent is the preferred density for IPUMS-International census samples. As noted above, of the 144 census samples in the database that could be drawn to IPUMS specifications, 127 attain this level of precision. A ten percent density is sufficient for at least rudimentary statistical analysis of geographical units with as few as 20,000 inhabitants.

Table 2 displays the implications of sample density for geographical analysis of the 4,378 towns in the 2011 Census of India. A ten percent sample would make it possible to set the threshold for identifying towns at 20,000. Such a threshold would mean that 75% of towns and 91% of town residents could be analyzed by town name in a statistically meaningful way. Class IV-VI towns would not be identified and therefore not analyzable by name. A sample density of five percent would raise the threshold to 100,000 inhabitants, suppress the names of Class III and II towns, and reduce the number of analyzable towns to a mere 9%, but account, nonetheless, for slightly over two-thirds of the population resident in towns. A 1% sample would require raising the threshold to 200,000 inhabitants, suppressing 96% of town names, and reducing the analysis to 42% of town populations. At 0.1%, the sample would still consist of over a million person records, but only 45 "towns" (urban agglomerations) would be analyzable by name, and spatial analysis would be severely limited due to the small sample size. At the other extreme, with a sample density of twenty percent, towns of 10,000 or more would be analyzable by name, accounting for 75% of towns and 97% of town residents. However, a sample density of 20% would require a heavy handed anonymization of other census questions to attain the necessary level of confidentiality.

Table 2. Census sample density and statistical analysis of urban populations
Example: 4,378 Towns 2011 Census of India (total population = 286,119,689)

		Towns			
Class	Pop.	N	% of N	Population	Sample density
IV-VI	<20k	1,073	75%	90.6%	10%
III	<50k	3,585	18%	78.4%	8.5%
II	<100k	3,992	9%	68.6%	5%
I*	<200k	4,213	4%	58.2%	1%

Source: http://www.censusindia.gov.in/towns/town.html, Office of the Registrar General.

Note: Official Population size-classes are: Class I: 100,000 and above; Class II: 50,000 to 99,999; Class III: 20,000 to 49,999; Class IV: 10,000 to 19,999; Class V: 5,000 to 9,999 and Class VI: Less than 5,000 persons.

*Here, for purposes of illustration, we have reduced Class I to 100,000-199,999.

Note that regardless of the threshold all classes of towns are represented in the sample. Only the number identified and therefore susceptible to analysis by name would vary. The basic rule is that the higher the density, the lower the threshold for identifying geographic units by name, the finer the grain of spatial analysis, and the coarser the grain of social, economic and demographic analysis. Most statistical offices participating in the IPUMS-International project agree that the ten percent density is the optimum for minimizing confidentiality risks and maximizing spatial and social detail.

High precision samples are cumbersome unless there is a means of disseminating microdata in a readily useable way. The IPUMS extraction system is designed precisely to manage such vast quantities of microdata to facilitate the dissemination of high precision datasets with fine geographic and social detail. As a test for this paper, we submitted an extract request for the 2000 and 2010 censuses of Indonesia—both with 10% sample densities (20,112,539 and 22,928,795 unit records, respectively). We choose to make an extract for a single district, the Regency of Bangli in the province of Bali, with 19,158 and 21,225 person records. Less than four minutes after submission, the extract was constructed and ready for download.

The experiment confirms that extraction time is not a constraint with IPUMS samples and that even the smallest identified geographic unit may be easily extracted for analysis. The IPUMS extract feature is ideal for mega-countries such as India with populations on a continental-scale. To see how the IPUMS extraction system works, we have posted a tutorial at: http://www.hist.umn.edu/~rmccaa/IPUMSI/ipumsi_small_area_example.ppt.

In conclusion, the matter is not that Indian census microdata are inadequate. Rather it is that they are inaccessible from a 21st century web-based platform. The lack of a time series of Indian census samples in the IPUMS-International database is a serious limitation not only for Indian social science and health research but also for understanding the great social, economic and demographic transformations of recent decades. Then too, there is a complete absence of access to Indian census microdata by researchers at the World Bank, OECD, UN Population Division and other international agencies as well as by Indian policy planners. We invite the ORGI to entrust ten percent household samples for the entire series of extant Indian census microdata for integration into the IPUMS-International database to open access to researchers in India and around the world.

References.

Alexander, J.T.; Davern, M.; and Stevenson, B. 2010. "Inaccurate Age and Sex Data in the [United States] Census PUMS Files: Evidence and Implications," *Public Opinion Quarterly*, 10 (Aug 10), pp. 1-10. doi: 10.1093/poq/nfq033

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

Esteve, A., J. Garcia and R. McCaa. 2011. "Comparative perspectives on Marriage and International Migration, 1970-2000: findings from IPUMS-International census microdata samples," *Seminar on Global Perspectives on Marriage and International Migration*, Seoul, South Korea: IUSSP Scientific Panel, Oct. 20-21.

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

McCaa, R., S. Ruggles and M. Sobek. 2010. "preparation,". in J. Domingo-Ferrer and E. Magkos (Eds.): *Privacy in Statistical Data 2010*, LNCS 6344. Springer, Heidelberg, pp.74-84.

Minnesota Population Center. 2012. *Integrated Public Use Microdata Series – International: Version 6.1*. Minneapolis: University of Minnesota: https://www.ipums.org/international.

[Parker, J.] 2011. "Herr and Madame, Señor and Mrs. Research at last begins to cast some light on the extent, causes and consequences of cross-border marriages," *Economist*, Nov. 12. http://www.economist.com/node/21538103

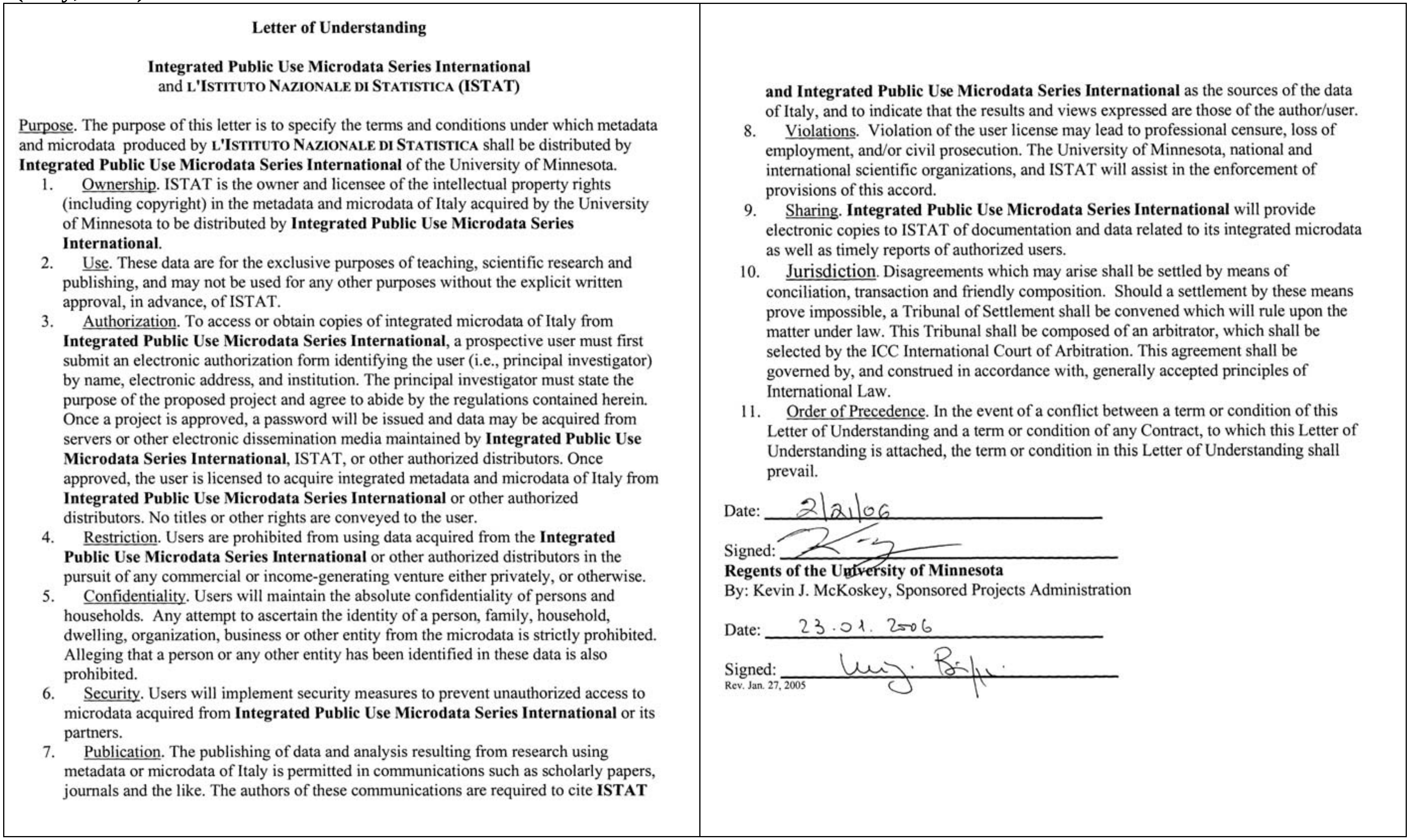
Premi, M.K. 2001. "Characteristics of the Population Census," in C.P. Chandrasekar and Jendhyala B.G. Tilsk (eds.), India's Socio-economic Database: Surveys of Selected Areas. Indian Council of Social Science Research, New Delhi, pp. 287-357.

Sobek, M. and S. Kennedy. 2009 The development of family interrelationship variables for international census data, Minnesota Population Center. https://international.ipums.org/international/resources/misc_docs/pointer_working_paper_2009.pdf .

Trewin, D. 2007. "A Review of IPUMS-International." Unpub. http://www.hist.umn.edu/~rmccaa/IPUMSI/trewin_ipums_report.pdf

Appendix A.

Example of Uniform Memorandum of Understanding between the University of Minnesota and National Statistical Offices (Italy, 2006).



Letter of Understanding

Integrated Public Use Microdata Series International
and **L'ISTITUTO NAZIONALE DI STATISTICA (ISTAT)**

Purpose. The purpose of this letter is to specify the terms and conditions under which metadata and microdata produced by **L'ISTITUTO NAZIONALE DI STATISTICA** shall be distributed by **Integrated Public Use Microdata Series International** of the University of Minnesota.

1. Ownership. ISTAT is the owner and licensee of the intellectual property rights (including copyright) in the metadata and microdata of Italy acquired by the University of Minnesota to be distributed by **Integrated Public Use Microdata Series International**.
2. Use. These data are for the exclusive purposes of teaching, scientific research and publishing, and may not be used for any other purposes without the explicit written approval, in advance, of ISTAT.
3. Authorization. To access or obtain copies of integrated microdata of Italy from **Integrated Public Use Microdata Series International**, a prospective user must first submit an electronic authorization form identifying the user (i.e., principal investigator) by name, electronic address, and institution. The principal investigator must state the purpose of the proposed project and agree to abide by the regulations contained herein. Once a project is approved, a password will be issued and data may be acquired from servers or other electronic dissemination media maintained by **Integrated Public Use Microdata Series International**, ISTAT, or other authorized distributors. Once approved, the user is licensed to acquire integrated metadata and microdata of Italy from **Integrated Public Use Microdata Series International** or other authorized distributors. No titles or other rights are conveyed to the user.
4. Restriction. Users are prohibited from using data acquired from the **Integrated Public Use Microdata Series International** or other authorized distributors in the pursuit of any commercial or income-generating venture either privately, or otherwise.
5. Confidentiality. Users will maintain the absolute confidentiality of persons and households. Any attempt to ascertain the identity of a person, family, household, dwelling, organization, business or other entity from the microdata is strictly prohibited. Alleging that a person or any other entity has been identified in these data is also prohibited.
6. Security. Users will implement security measures to prevent unauthorized access to microdata acquired from **Integrated Public Use Microdata Series International** or its partners.
7. Publication. The publishing of data and analysis resulting from research using metadata or microdata of Italy is permitted in communications such as scholarly papers, journals and the like. The authors of these communications are required to cite **ISTAT and Integrated Public Use Microdata Series International** as the sources of the data of Italy, and to indicate that the results and views expressed are those of the author/user.
8. Violations. Violation of the user license may lead to professional censure, loss of employment, and/or civil prosecution. The University of Minnesota, national and international scientific organizations, and ISTAT will assist in the enforcement of provisions of this accord.
9. Sharing. **Integrated Public Use Microdata Series International** will provide electronic copies to ISTAT of documentation and data related to its integrated microdata as well as timely reports of authorized users.
10. Jurisdiction. Disagreements which may arise shall be settled by means of conciliation, transaction and friendly composition. Should a settlement by these means prove impossible, a Tribunal of Settlement shall be convened which will rule upon the matter under law. This Tribunal shall be composed of an arbitrator, which shall be selected by the ICC International Court of Arbitration. This agreement shall be governed by, and construed in accordance with, generally accepted principles of International Law.
11. Order of Precedence. In the event of a conflict between a term or condition of this Letter of Understanding and a term or condition of any Contract, to which this Letter of Understanding is attached, the term or condition in this Letter of Understanding shall prevail.

Date: 2/21/06

Signed:

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

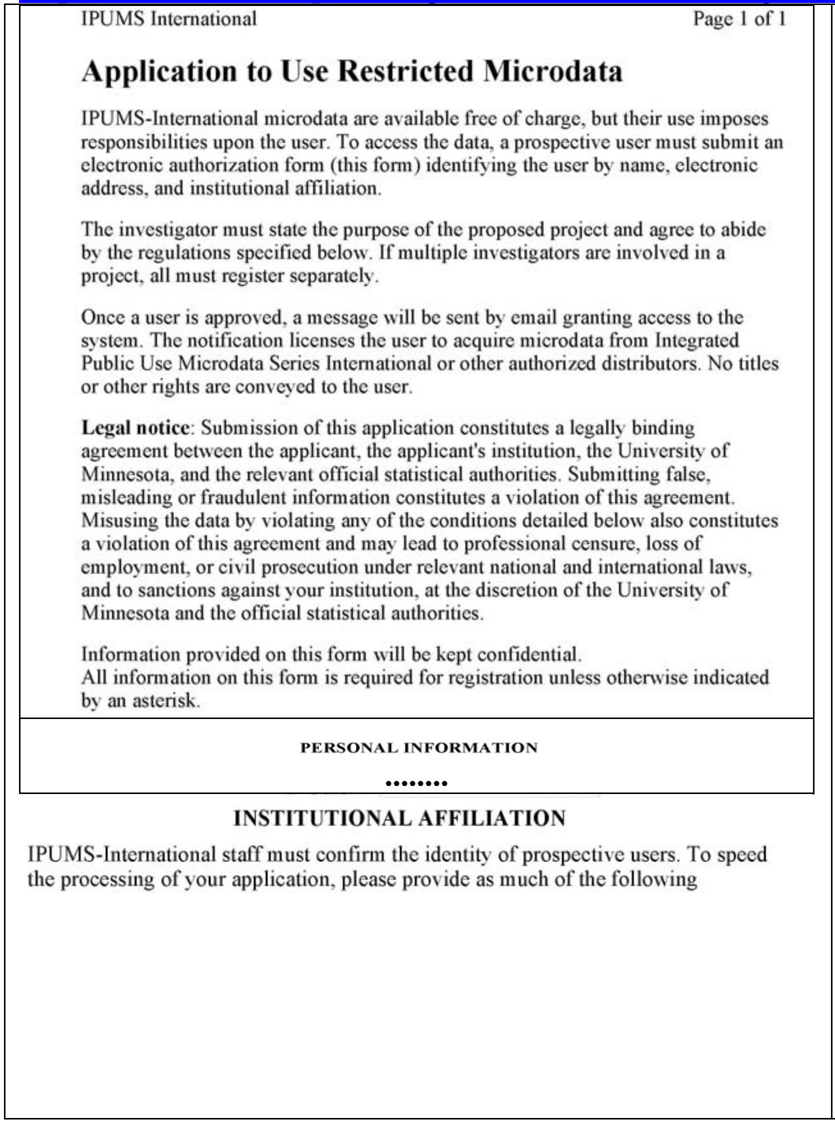
Date: 23.01.2006

Signed:

Rev. Jan. 27, 2005

Appendix B.

Snippets of Application Form to Use Restricted Microdata disseminated by IPUMS-International. See: https://international.ipums.org/international-action/register/0



IPUMS International Page 1 of 1

Application to Use Restricted Microdata

IPUMS-International microdata are available free of charge, but their use imposes responsibilities upon the user. To access the data, a prospective user must submit an electronic authorization form (this form) identifying the user by name, electronic address, and institutional affiliation.

The investigator must state the purpose of the proposed project and agree to abide by the regulations specified below. If multiple investigators are involved in a project, all must register separately.

Once a user is approved, a message will be sent by email granting access to the system. The notification licenses the user to acquire microdata from Integrated Public Use Microdata Series International or other authorized distributors. No titles or other rights are conveyed to the user.

Legal notice: Submission of this application constitutes a legally binding agreement between the applicant, the applicant's institution, the University of Minnesota, and the relevant official statistical authorities. Submitting false, misleading or fraudulent information constitutes a violation of this agreement. Misusing the data by violating any of the conditions detailed below also constitutes a violation of this agreement and may lead to professional censure, loss of employment, or civil prosecution under relevant national and international laws, and to sanctions against your institution, at the discretion of the University of Minnesota and the official statistical authorities.

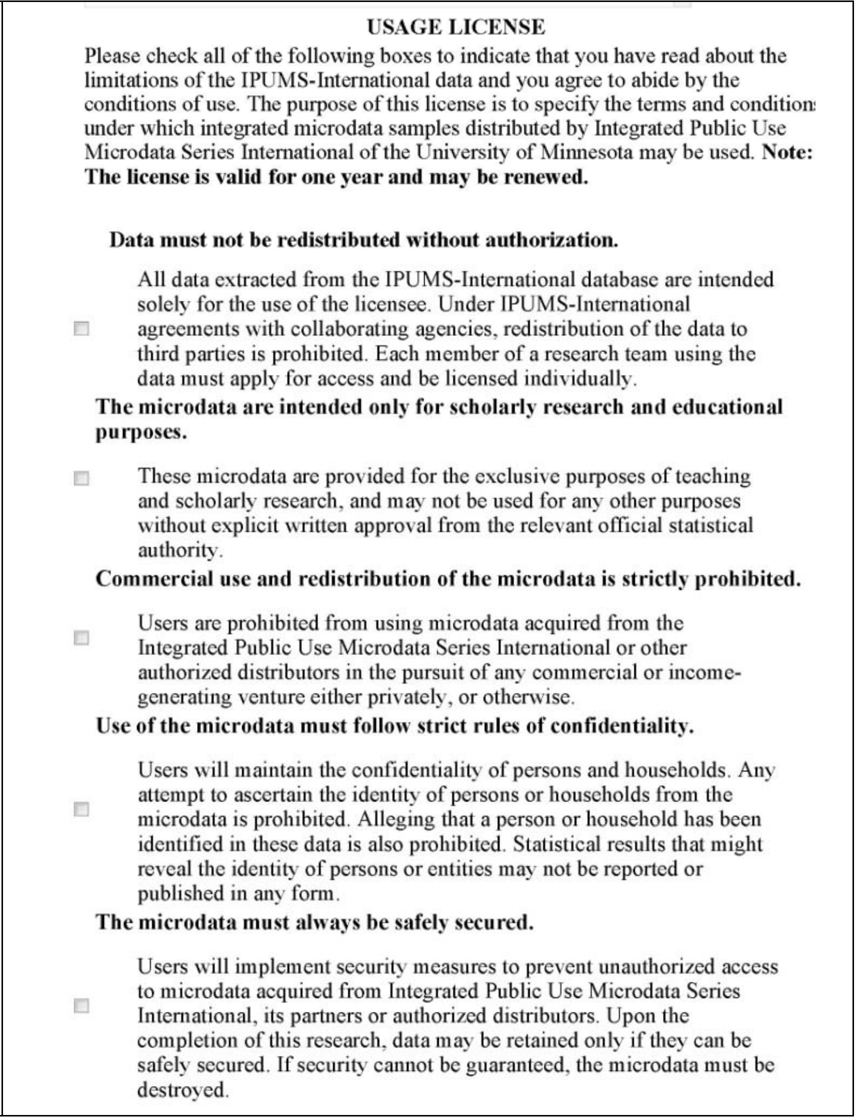
Information provided on this form will be kept confidential.
All information on this form is required for registration unless otherwise indicated by an asterisk.

PERSONAL INFORMATION

……..

INSTITUTIONAL AFFILIATION

IPUMS-International staff must confirm the identity of prospective users. To speed the processing of your application, please provide as much of the following



USAGE LICENSE

Please check all of the following boxes to indicate that you have read about the limitations of the IPUMS-International data and you agree to abide by the conditions of use. The purpose of this license is to specify the terms and condition under which integrated microdata samples distributed by Integrated Public Use Microdata Series International of the University of Minnesota may be used. **Note: The license is valid for one year and may be renewed.**

Data must not be redistributed without authorization.

□ All data extracted from the IPUMS-International database are intended solely for the use of the licensee. Under IPUMS-International agreements with collaborating agencies, redistribution of the data to third parties is prohibited. Each member of a research team using the data must apply for access and be licensed individually.

The microdata are intended only for scholarly research and educational purposes.

□ These microdata are provided for the exclusive purposes of teaching and scholarly research, and may not be used for any other purposes without explicit written approval from the relevant official statistical authority.

Commercial use and redistribution of the microdata is strictly prohibited.

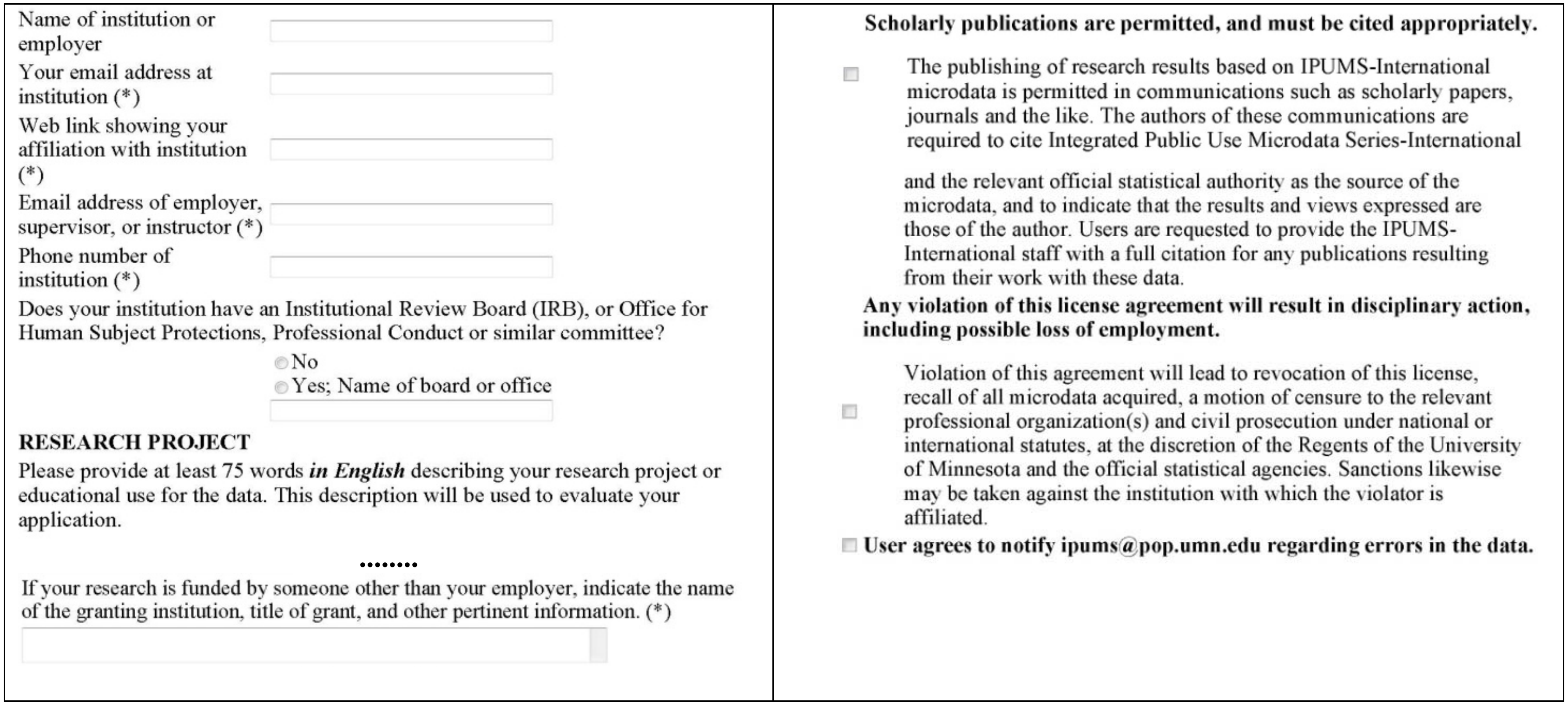
□ Users are prohibited from using microdata acquired from the Integrated Public Use Microdata Series International or other authorized distributors in the pursuit of any commercial or income-generating venture either privately, or otherwise.

Use of the microdata must follow strict rules of confidentiality.

□ Users will maintain the confidentiality of persons and households. Any attempt to ascertain the identity of persons or households from the microdata is prohibited. Alleging that a person or household has been identified in these data is also prohibited. Statistical results that might reveal the identity of persons or entities may not be reported or published in any form.

The microdata must always be safely secured.

□ Users will implement security measures to prevent unauthorized access to microdata acquired from Integrated Public Use Microdata Series International, its partners or authorized distributors. Upon the completion of this research, data may be retained only if they can be safely secured. If security cannot be guaranteed, the microdata must be destroyed.



Name of institution or employer	
Your email address at institution (*)	
Web link showing your affiliation with institution (*)	
Email address of employer, supervisor, or instructor (*)	
Phone number of institution (*)	

Does your institution have an Institutional Review Board (IRB), or Office for Human Subject Protections, Professional Conduct or similar committee?

○ No

○ Yes; Name of board or office

RESEARCH PROJECT

Please provide at least 75 words ***in English*** describing your research project or educational use for the data. This description will be used to evaluate your application.

••••••••

If your research is funded by someone other than your employer, indicate the name of the granting institution, title of grant, and other pertinent information. (*)

Scholarly publications are permitted, and must be cited appropriately.

□ The publishing of research results based on IPUMS-International microdata is permitted in communications such as scholarly papers, journals and the like. The authors of these communications are required to cite Integrated Public Use Microdata Series-International and the relevant official statistical authority as the source of the microdata, and to indicate that the results and views expressed are those of the author. Users are requested to provide the IPUMS-International staff with a full citation for any publications resulting from their work with these data.

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

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

□ **User agrees to notify ipums@pop.umn.edu regarding errors in the data.**

Appendix C.
Educational attainment: IPUMS integrated codes for most recent census sample: 11 Asian countries

IPUMS	Label	Cambodia	China	India	Indonesia	Iran	Malaysia	Nepal	Pakistan	Philippines	Thailand	Vietnam
Code		2008	1990	2004	2010	2006	2000	2001	1998	2000	2000	2009
0	NIU (not in universe)	136,274	1,418,185	·	2,253,453	131,235	·	376,202	1,944,463	935,577	43,640	1,517,591
100	LESS THAN PRIMARY COMPLETED	·			·	·	·	·	·	·	·	·
110	No schooling	297,550	2,145,035	220,227	1,986,754	41,776	100,909	1,118,000	6,375,658	466,783	55,479	892,633
120	Some primary	468,764	2,238,032	96,159	4,131,163	125,249	122,425	373,487	1,525,254	1,665,337	229,206	2,643,510
130	Primary (4 years)	·	·	·	·	·	·	·	·	·	·	·
	PRIMARY COMPLETED, LESS THAN SECONDARY											
	Primary completed											
211	Primary (5 years)	·	·	88,352	·	262,244	·	274,958	·	·	·	3,810,866
212	Primary (6 years)	256,570	2,822,479	·	6,539,863	·	80,005	·	2,038,363	1,967,457	116,450	·
	Lower secondary completed											
221	General and unspecified track	119,439	2,247,161	84,369	3,595,440	234,096	86,632	140,342	683,765	·	62,897	3,468,942
222	Technical track	·	·	·	·	·	·	·	·	·	·	·
	SECONDARY COMPLETED											
	General or unspecified track											
311	General track completed	41,385	640,916	49,669	3,592,138	127,008	8,878	247,725	260,127	814,182	24,371	1,074,774
312	Some college/university	·	43,450	29,237	·	29,805	·	·	·	715,722	17,237	151,141
320	Technical track	·	·	·	·	·	·	·	·	·	·	·
321	Secondary technical degree	2,228	148,554	·	400,543	38,242	·	·	·	·	13,106	90,359
322	Post-secondary technical	4,224	82,642	6,117	401,387	·	1,608	598	15,892	159,614	14,991	·
400	UNIVERSITY COMPLETED	13,010	49,493	28,290	702,308	59,970	25,456	38,416	236,278	305,054	20,933	527,774
999	UNKNOWN/MISSING	677	·	413	·	250,200	9,387	13,517	22,224	388,084	6,209	·

Note: Case counts indicate the absolute number of persons in the sample with the corresponding code

Appendix D.
Person variables in Censuses of India compared with integrated samples for selected Asian countries

		Country	India (from form)			Cambodia	China	India	Indonesia	Iran	Malaysia	Nepal	Pakistan	Philippines	Thailand	Vietnam
IPUMS		Year	2011	2001	1991	2008	1990	NSSO2004	2010	2006	2000	2001	1998	2000	2000	2009
Mnemonic	**Label**	**Total**	**29**	**27**	**26**	**28**	**21**	**16**	**28**	**28**	**28**	**31**	**9**	**24**	**27**	**25**
RELATE	Relationship to hh head		X	X	X	X	X	X	X	X	X	X	X	X	X	X
AGE	Age		X	X	X	X	X	X	X	X	X	X	X	X	X	X
SEX	Sex		X	X	X	X	X	X	X	X	X	X	X	X	X	X
MARST	Marital status		X	X	X	X	X	X	X	X	X	X	X	X	X	X
EDATTAN	Educational attainment, intl. recode		X	X	X	X	X	X	X	X	X	X	X	X	X	X
LIT	Literacy		X	X	X	X	X	X	X	X	X	X	X	X	X	X
OCC	Occupation, unrecoded		X	X	X	X	X	X	.	X	X	X	.	X	X	X
INDGEN	Industry, general recode		X	X	X	X	X	X	X	X	X	X	.	X	X	X
IND	Industry, unrecoded		X	X	X	X	X	X	X	X	X	X	.	X	X	X
OCCISCO	Occupation, ISCO general, 1-digit		X	X	X	X	X	X	.	X	X	X	.	X	X	X
CLASSWK	Class of worker		X	X	X	X	.	X	X	X	X	X	.	X	X	X
SCHOOL	School attendance		X	X	X	X	.	X	X	X	X	X	.	X	X	X
EMPSTAT	Employment status		X	X	X	X	X	X	X	X	X	X	.	.	.	X
NATIVTY	Nativity status		X	X	X	X	.	.	X	X	X	X	.	X	X	.
CHBORN	Children ever born		X	X	X	X	X	.	X	X	.	X	.	.	X	X
CITIZEN	Citizenship					.	.	.	X	X	X	X	X	.	X	.
CHSURV	Children surviving		X	X	X	X	X	.	X	X	.	X	.	.	X	X
BPLCTRY	Country of birth		X	X	X	X	.	.	X	.	X	X	.	X	X	.
	Administrative area of birth		X	X	X	X	.	.	X	.	X	X	.		X	.
RELIG	Religion		X	X	X	X	.	X	X	X	X	X	X	X	X	.
BIRTHYR	Year of birth		X			.	X	.	X	X	.	.	.	.	X	X
BIRTHMO	Month of birth		X			.	X	.	X	X	.	.	.	.	X	X
NATION	Country of citizenship					.	.	.	X	X	X	X	.	.	X	.
EMPSECT	Sector of employment		X	X	X	X	.	X	.	.	X	.	.	X	.	X
MGYRS1	Years residing in current locality		X	X	X	X	.	.	.	X	.	X	.	.	X	.
DISABLE	Disability status		X	X		X	.	.	X	X	X	X	.	X	.	X
DISEMP	Employment disability					.	X	X	.	.	X	X	.	.	X	X
DISBLND	Blind or vision-impaired					X	.	.	X	X	X	X	.	X	.	X
BRTHLYR	Number of births last year		X	X	X	X	X	.	.	X	.	X	.	.	X	.
ISCO88A	Occupation, ISCO-1988, 3-digit		X	X	X	X	.	.	.	X	X	.	.	X	X	.
MGRATE5	Migration status, 5 years					.	X	.	X	.	X	X	.	X	.	X
DISDEAF	Deaf or hearing-impaired					X	.	.	X	X	X	X	.	X	.	X
	Place of residence 5/10 years ago (national definition						X		X		X	X		X		X
	Ethnicity (national definition)		X	X	X		X		X		X	X		X		X
	Mother tongue (national definition)		X	X	X	X			X			X	X	X		
MGCAUSE	Reason for migration		X	X	X	X	X	.	.	X	.	.	.	.	X	.

Additional variables unique to Indian censuses					**2011**	**2001**	**1991**		**2011**	**2001**	**1991**
Scheduled Caste or Scheduled Tribe	X	X	X	Place of Birth code	X	X	X	Duration of residence	X	X	X
First other language known	X	X	X	POB rural/urban	X	X	X	Travel to work - distance	X	X	
Second other language known	X	X	X	POB district	X	X	X	mode of travel to work	X	X	
Seeking/available for work	X	X	X	POB state/country	X	X	X	Non-economic activity	X		
Have you ever worked before	X	X	X	Place of last residence	X	X	X				
Ex-Serviceman			X	PLR rural/urban	X	X	X				
Pensioner			X	PLR district	X	X	X				
				PLR state/country	X	X	X				