IPUMS-International: lessons from 10 years of archiving and disseminating census microdata

McCaa, Robert Thomas, Wendy L.

Minnesota Population Center, 50 Willey Hall, University of Minnesota
Minneapolis, MN 55455 USA

rmccaa@umn.edu wlt@pop.umn.edu

Introduction

The IPUMS-International project (www.ipums.org/international) has four main goals:

1. Inventory the currently existing microdata for the world's population censuses
2. Recover and preserve those at-risk
3. Construct an integrated database system containing both harmonized and non-harmonized microdata and metadata concepts, questions, and variables of population censuses of nations worldwide
4. Disseminate these data and their related metadata to researchers and policy makers at no cost.

Begun in 1999 with a five year grant from the National Science Foundation of the United States, the initiative now includes regional projects in Latin America (2003-2014), Europe (2004-2009) and Eurasia (2009-2014) funded by the National Institutes of Health. A fourth regional project is in preparation in Africa. During the initial ten years, the project Memorandum of Understanding has been endorsed by over 80 official statistical agencies encompassing over three-fourths of the world's population. The Memoranda are uniform in content—covering principles of ownership, access, security, and confidentiality protections—and have resulted in the launch of 130 integrated samples, encompassing 44 countries and totaling 279 million person records. An additional 124 datasets have been entrusted to the Minnesota Population Center (MPC) for incorporation into the IPUMS database over the next five years.

The first lesson that the IPUMS team has learned is that many national statistical offices are eager to disseminate census microdata but are hesitant to do so because of the difficult legal, administrative, technical and resource obstacles. The IPUMS initiative offers a proven, secure, reliable and virtually cost-free solution to these problems. The University of Minnesota assumes the legal and administrative responsibilities for licensing the microdata from each national statistical office and for enforcing a uniform dissemination agreement between the University, on the one hand, and researchers and their institutions, on the other. To solve technical problems, the IPUMS project has developed highly efficient, semi-automated tools to anonymize, integrate and disseminate the microdata. To integrate the metadata, source documentation is fed into an XML database which instantaneously generates dynamic metadata specific to any census question for any combination of countries and years. This greatly facilitates not only the integration of the microdata by the MPC team but also assists comprehension by researchers of subtle differences in concepts from census to census.

Archiving and preserving international census microdata and metadata

The IPUMS team is lead by population historians. Our first goal is to inventory and preserve census documentation and machine-readable microdata. In 2000, the first edition of our census microdata inventory was published in Handbook of International Census Microdata for Population Research, edited by P. Kelly Hall, R. McCaa, and G. Thorvaldsen. This has since been updated and a revised inventory can be accessed

here (www.hist.umn.edu/~rmccaa/IPUMSI/ipumsi_microdata_inventory.htm).

Census microdata exist in a wide variety of formats. While many of the contemporary censuses are well preserved, many older census files have fallen victim to time, lack of preservation funding, external political situations, or natural disaster. Bangladesh is an example of one of our most complex, and costly, undertakings to date. The commercial firm, Muller Media, working under contract with the IPUMS project, equipped a data recovery laboratory in the Bangladesh Bureau of Statistics and trained BBS staff to recover data from old main-frame tapes. Approximately 300 9-track tapes containing the 1981 census microdata were recovered, including a 1% sample with geographical coverage of 99.7% of the entire country. In addition 100% microdata was recovered for three-fourths of the enumerated population. The BBS is continuing the operation to restore data from some 7,000 tapes. In the case of Mali, old Benoulli boxes yielded 93.6% of the person records for the 1976 census. For the 1977 census of Romania, the percentage was 93.1. The Federal Statistical Office of Germany recovered 100% of the microdata for the 1971 census of the German Democratic Republic. For the censuses of Mongolia 1989 and Fiji 1976, IPUMS contributed to the re-keying of microdata from archived questionnaires, where the original microdata had been lost. The National Statistical Office of Mongolia is now entertaining a proposal to extend the recovery effort to the 1956 census. Morocco is considering a similar undertaking for the censuses of 1961 and 1971. The costs of these recovery efforts are trivial compared to the huge costs of the original census operations.

These and other smaller recovery projects successfully dealt with old storage media, obsolete data compression algorithms, physical damage, and loss of data. All data recovered through these processes, are migrated to an archive format, reviewed, checked for data quality and internal inconsistencies, documented regarding processing steps, and repatriated to the national statistical agency of origin.

The second lesson that we learned is that National Statistical Offices are eager to recover the statistical heritage represented by census microdata, but often lack the resources to do so. IPUMS technical assistance is often successful in restoring data from tapes and other obsolete media even for those stored for long periods of time in less than ideal conditions.

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. Users have access to the reformatted content, original documents, and ancillary documentation. The system uses a metadata-centric approach, allowing the research staff to manipulate simple but highly-structured documents to drive both the data processing and the web software. A unique XML markup identifies all elements necessary to guide the recoding and documentation of variables and to associate each variable with its relevant enumeration materials. The data, documentation, and dissemination software systems are all driven by the same metadata, which ensures that they always remain synchronized.

The second area of archiving activity is the cataloging and preservation of over 25,000 documents related to international censuses covering a period of approximately 1920 to date. 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 documentations from the United States Census Bureau International Collection, plus contributions of materials from 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). In 2001, the United Nations Statistics Division entrusted its archive of historical census documentation, including enumeration forms for most countries dating from the 1980s and earlier, to the Minnesota Population Center. A collection of scanned enumeration forms, consisting of the 1960-1990 census rounds, "World Census Questionnaires," was published by the MPC and is available upon request.

In addition the MPC has produced individual country collections for over 40 national statistical offices, providing full scans of critical documents and a listing with first page scans of our full collection for each country. These products allow countries to easily identify any documents not present in the IPUMS collection as well as to provide quality scanned copies of their census related documents. In a number of cases, the documents repatriated in this manner have been missing from the collections of the national statistical organization and have been especially appreciated. As a companion to the "World Census Questionnaire" we have published volumes of "Census Enumerator Manuals" for both Latin American and Africa. The "IPUMS-Latin America: Census Enumerator Manuals" is available on-line in the University of Minnesota Digital Conservancy. <http://conservancy.umn.edu/handle/5948>. A copy of "IPUMS-Africa: Census Enumerator Manuals and Forms" has been produced for distribution at ISI2009. 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, and lack of preservation. Scanning preserves the content and provides a means of access and dissemination for both the MPC and the National Statistical Offices of the originating countries.

Full scans are being produced for all documents critical to processing census samples in the IPUMS collection, all enumeration forms and enumerator manuals, and the collection from the United States Census Bureau (to replace the physical collection provided to the MPC). All other documents have their first page scanned and a bibliographic record created. Full scans will be produced as demand and funding dictate.

Constructing an integrated microdata and metadata system

IPUMS is *not* simply a conduit for passing census samples from National Statistical Offices along to researchers. Instead, typically, two or more years of labor are invested by the IPUMS project in preparing anonymized, integrated 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. There are five steps to the IPUMS process before microdata are disseminated.

1. Confirm the integrity and validity of the source microdata and metadata
2. Draw and anonymize the high precision sample on which all subsequent work is based
3. Integrate the microdata
4. Integrate the metadata
5. Confirm the integrity and validity of the integrated microdata sample and metadata

Steps one and two are conducted on the original source microdata entrusted to the Minnesota Population Center. These microdata are never disseminated to anyone or any institution—other than the

corresponding National Statistical Office-owner. For this reason access to these data is restricted to senior civil service staff at the MPC who are thoroughly trained in protecting data security. These data are exceedingly sensitive and for that reason only seasoned, specially trained, full time researchers with a need to complete the first two tasks of the IPUMS process have access to these data. MPC employees are subject to civil fine (up to US$250,000) and criminal prosecution for violation of security procedures. The University legal authority assumes responsibility for protecting the total confidentiality of these datasets. A complete review of these processes was conducted on-site by Mr. Dennis Trewin, the chairman of the UN-ECE joint-committee on Statistical Confidentiality and Microdata Access and President ex-officio of the International Statistical Institute. Mr. Trewin's report concludes:

> "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. … The security of the computing environment used by IPUMS-International is first class and appears to be of the standard of the best statistical offices."

Consider each of the five steps of the IPUMS integration process in more detail.

1. Confirm integrity and validity. The microdata are exhaustively evaluated by IPUMS senior staff to resolve issues of data integrity and validity. Note, however, to date the project does not perform data editing or imputation. Instead effort is focused on ensuring the household structure of records and confirming that sample statistics approximate official published figures. Since the purpose of the sample is to provide a dataset for analysis, there is no need to insure that samples replicate published census results to the last digit.

2. Draw and anonymize the sample. One of the most important stratifying variables in survey research and in drawing high precision census microdata samples is geography. Geography is related to a great number of variables researchers are interested in studying and therefore increases the efficiency of stratified samples. Many of the IPUMS-International samples capitalize on *implicit* geographic stratification. The raw census files used to construct IPUMS samples are typically geographically organized within districts. Systematic random samples of the censuses capitalize on this low-level geographic sorting. By ensuring a representative geographic distribution of sampled cases, they are equivalent to extremely fine geographic stratification with proportional weighting. Since many economic and demographic characteristics are highly correlated with geographic location, this implicit stratification yields substantially greater precision than would a simple random sample of households. As part of the IPUMS project, we are developing stratification variables that allow researchers to make reliable variance estimates from implicitly stratified samples.

Almost all the statistical agency partners of the IPUMS project have endorsed the use of implicitly stratified samples of households. Thirty-seven National Statistical Offices have entrusted complete sets of census microdata to facilitate the drawing of implicitly stratified samples by the MPC. In Europe, almost all the statistical agencies have drawn new samples using IPUMS specifications. IPUMS sample densities typically range between 5 and 10%. Lower densities are provided by countries where privacy matters are a greater issue than quality (Netherlands, United Kingdom) or, as in the case of 1960 round of censuses, where only low precision samples survive.

In cases where fully anonymized samples are entrusted to the project, no further statistical confidentiality measures are imposed. However, in many cases, full records are provided to the project, including detail sufficient to pose a theoretical risk of re-identification. To minimize risk, statistical confidentiality edits are performed by the IPUMS project. The lowest level of geography to be released is identified (e.g., for European countries, typically "NUTS3") and all finer geographic variables are suppressed. Any technical variables that could be used to identify records or which have been identified as sensitive within the original data are also suppressed. Variables with very small population categories are recoded into larger groups (e.g., grouping a detailed occupation with its parent category) and top- or bottom-coding is performed where needed (e.g., income). Finally, the sequence of dwellings within the smallest geographic unit identified in the data is randomized, so geography cannot be inferred. An undisclosed fraction of cases is randomly swapped across geographic districts to add uncertainty about the origin of any particular record. Finally, a new serial number is generated to reflect the ordering of the file.

3. Microdata Integration. The principal benefit of IPUMS to researchers and NSOs alike is integration—integration of both microdata and metadata. For decades, many NSOs have provided census samples for academic and policy research, but few statistical offices re-examine earlier samples to harmonize successive datasets or to draft new documentation to facilitate comparative analysis of two or more censuses. At best, as soon as the final data cleaning is complete, the more modern statistical offices construct a census sample and a data dictionary for researchers, Five or ten years later, with the ensuing census, the process is repeated with little guidance on enhancing the comparability of successive census datasets.

We must reiterate that the IPUMS project does *not* disseminate census files entrusted by national statistical offices. Instead high-precision census samples are anonymized (McCaa et. al. 2006) and 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 by means of meticulous study of comprehensive original source documentation and after extensive analysis of the microdata. Thousands of hours are devoted to analyze, discuss, debate, draft, 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.

As an example of the IPUMS method of integrating a variable, consider the concept "married". In recent decades, as the United Nations Statistics Division Principles and Recommendations have evolved, there is increasing precision in definitions. Nonetheless, in practice, some censuses simply refer to "married," while others report "formally married" according to civil law or religious convention or both. Still other censuses distinguish informal unions from formal ones. The IPUMS system seeks to retain all significant distinctions in the original microdata. Thus for "married otherwise undefined", the code is "200". Formally married is "210". Formal civil marriage is "211", which contrasts with religious marriage "212," both religious and civil "213", either "214", traditional "215", and polygamous "217". Consensual unions are coded "220". For complete details, see the IPUMS metadata for "Marital Status": https://international.ipums.org/international-action/codes.do?mnemonic=MARST Successful international integration must document these distinctions so that researchers may readily be informed of these and thousands of other details. As the next section indicates, the IPUMS integrated metadata describe these general details as well as subtle distinctions for specific countries and individual censuses.

IPUMS integration has enjoyed such success that some statistical agency partners prefer to use the integrated IPUMS microdata rather than their own original source data. For example, DANE-Colombia, the first statistical agency to participate in the IPUMS collaboratory, is using the five integrated Colombian census samples (12.3 million person records) to construct a nationally integrated dataset with metadata in the Spanish. With IPUMS assistance, DANE is simplifying internationally integrated datasets to a national system. It is more efficient to simplify an international integration because where a national integration is performed first; important details may be unwittingly sacrificed as trivial at the national level but are essential for successful international integration.

4. Metadata integration. Metadata integration is essential if microdata integration is to succeed. Integrated metadata relieves researchers of the task of studying documentation *en toto* of every census for changes or deviations in concepts and definitions. Instead, before microdata are integrated into the IPUMS system, experts carefully consider all the documentation and analyze the microdata to write new, comprehensive documentation that spells out common practices and discusses significant differences and discrepancies. Once the data are released, researchers study the integrated metadata confident that their attention is directed to issues of greatest salience for the research questions at hand.

The IPUMS eXtensible Markup Language (XML) tool, by means of a few clicks, facilitates navigation of both source and integrated metadata in any way desired. For example, to compare the wording of the employment status variable, select the countries and census years desired, then click "employment status", and "enumeration text". This allows the researcher to compare the precise wording, in English, of the question on the form as well as the instructions to the enumerators for all selected census.

a. Censuses and samples. IPUMS metadata offer detailed 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, if any. Images of census enumeration forms and instructions manuals are available in the official language and the text in English translation. Each sample is described with regard to source, sample design, sampling unit, sample fraction, number of person records, sample weights, dwelling or housing units, vacant dwellings, households, group quarters and special populations.

b. Variable descriptions, source texts, and codes. IPUMS metadata define each integrated variable and describe basic characteristics: availability by census, universe of the variable or question, codes, source (enumeration) text, and non-harmonized variables used for integration. Access to this information is through clickable hypertext on the IPUMS website. A general comparability discussion is provided for every variable, with country or census specific discussions focusing on departures from standard practice. The purpose of these discussions is to highlight important contrasts. Clicking "Enumeration text" leads to source questions and corresponding instructions for each selected census in English. 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.

Coding of variables in the IPUMS system may be viewed in either general or detailed versions using one of two views. The first view is a table containing an "X" indicating the presence of the code in a specific census, while the second provides the exact, un-weighted case count in the integrated sample. The "codes" table is handy for determining whether specific attributes are present in sufficient quantity for the

contemplated research as well as planning recodes for specific analytical purposes.

5. Validation and Certification. Before samples are made available to researchers, the entire database is checked for consistency and accuracy. This requires verification of hundreds of thousands of coding decisions. The process is facilitated by the fact that the database contains both integrated and non-harmonized variables. Verification is performed by cross-tabulating each integrated variable by its corresponding non-harmonized version of the variable. Initially IPUMS-International focused on a number of harmonized variables that were common across nations and over time. As historians, comparisons over time as well as space are vitally important. However, mid-way through the project it became clear that there was a demand to retain both the content of and access to the unharmonized variables that are specific to each census sample. The importance of facilitating access to the unharmonized variables is two-fold. First, it provides the source material from which the harmonized variables are constructed, allowing researchers a fuller understanding of the harmonization process. Second it preserves the original census structure and content, providing a reflection of changes in a country's census series over time and differences between nations concerning what concepts are covered and how they are expressed in national censuses. In 2006 over 5000 unique sample-specific non-harmonized variables were added to the IPUMS database. Each variable has its universe documented and empirically verified. Although some variables are suppressed for confidentiality reasons or obvious data errors, the goal is to provide as complete a picture of the original census as possible.

To obtain an outside evaluation of the integration process performed by the IPUMS team, the National Statistical Office of Argentina (INDEC) was contracted to conduct an exhaustive analysis of the integration of samples of the Argentine censuses of 1970, 1980, 1991 and 2001. INDEC experts compared the frequencies for each variable and code against the original microdata and metadata entrusted to IPUMS. From the tens of thousands of words and codes of metadata, barely a handful of errors, misinterpretations or misunderstandings were discovered. All were considered minor. This important outside evaluation—accomplished on-site in INDEC's Buenos Aires offices without the presence of IPUMS personnel—attests to the trustworthiness of IPUMS integrations. What INDEC did can be done by any statistical agency working in the convenience of their own offices. The IPUMS database provides tools for the expert user to cross-check every integration decision made by the IPUMS team so that little doubt remains about the significance, quality, or transparency of the entire integration process.

The third lesson that we have learned is that integration is difficult, but the IPUMS system has streamlined the process so that it is possible to harmonize 15-20 censuses per year to a high degree of precision and to the satisfaction of National Statistical Office-owners of the microdata as well as for users—academic researchers and policy makers.

Disseminating microdata and metadata to accredited researchers world-wide

To date, more than 3,000 researchers have obtained approval to access the database, representing 76 countries, and hundreds of universities and international organizations. Note that researchers must first be approved before access to any microdata on the IPUMS-International website is permitted. Access is obtained by submitting a detailed application form and agreeing to each condition of use as required by the project Memorandum of Understanding. Once approved, microdata are provided in the form of extracts, custom tailored to each researcher's need. The IPUMS database is not distributed *en toto* and the ability to

reassemble subsets into a replication of the whole is effectively curtailed by the IPUMS dissemination system.

To request an extract, the researcher must first sign in by entering the registered password then make a series of selections by means of point-and-click menus, specifying country (or countries), census year(s), sample(s), variables and sub-populations, as well as metadata format (SAS, SPSS, or STATA are supported). Once the selections are complete, there is an opportunity to review or revise before final submission of the request. Then, once submitted, the IPUMS extract engine registers the request and places it in a data processing queue. When the extract is ready (usually in a matter of hours, if not minutes), the researcher is notified by email that the data should be retrieved within 72 hours. A link is provided to a password-protected page for downloading the specific extract via SSL (Secure Sockets Layer) protocol. Microdata are transmitted using the 128-bit encryption standard, matching the level used today by the banking and other industries where security and confidentiality is 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 provided in ASCII format. The IPUMS help-desk responds to user questions and queries and assembles copies of publications for transmission to the respective National Statistical Offices.

Conclusions

Thanks to widespread support by official statistical agency partners, IPUMS has already become one of the largest demographic databases in the world. National Statistical Offices not presently participating in the IPUMS initiative are respectfully invited to consider participation. Researchers who have a need to analyze census microdata are cordially invited to visit the website and use the data. The final lesson that we have learned is that there is a considerable demand for census microdata. Indeed, in the case of the United States, IPUMS-USA is the single most frequently cited data-source in the premier journal of population studies, *Demography.* If IPUMS-International is successful, it is likely to become one of the most widely used sources by academic researchers and policy makers requiring census samples for analysis.

REFERENCES

Conference of European Statisticians. (2007). *Managing Statistical Confidentiality and Microdata Access: Principles and Guidelines on Good Practice*. http://www.unesce.org/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf Geneva: United Nations Economic Commission for Europe.

Esteve, Albert and Matthew Sobek. (2003). Challenges and Methods of Census Harmonization. *Historical Methods* 36: 66-79.

McCaa, Robert, and Steven Ruggles (2000). "IPUMS-International: A Global Project to Preserve Machine-Readable Census Microdata and Make Them Usable." In *Handbook of International Historical Microdata*, ed. By Patricia Kelly Hall, Robert McCaa, and Gunnar Thorvaldsen, 335-346. Minneapolis, MN: Minnesota Population Center https://international.ipums.org/international/microdata_handbook.shtml.

McCaa, Robert, Steven Ruggles, Michael Davern, Tami Swenson, Krishna Mohan Palipudi. (2006). "IPUMS-International High Precision Population Census Microdata Samples: Balancing the Privacy-Quality Tradeoff by Means of Restricted Access Extracts," Privacy in Statistical Databases. Berlin: Springer, pp. 375-382.

Thomas, Wendy L. and Robert McCaa. “Archiving Census Documentation and Microdata: Preserving Memory, Increasing Stakeholders", *Notas de Población* XXIX:75 (2003), 303-320.

Trewin, Dennis. (2007). A review of IPUMS-International. Unpublished.

disseminate census microdata but are hesitant to do so because of the difficult legal, administrative, technical. and resource obstacles. The IPUMS initiative offers a proven, secure, reliable and virtually cost-free solution.Â The second area of archiving activity is the cataloging and preservation of over 25,000 documents related to international censuses covering a period of approximately 1920 to date. These documents provide a rich source of material that cover technical specifications, preparations for census activities, involvement of.Â â€œIPUMS-International: A Global Project to Preserve Machine-Readable Census Microdata and Make Them Usable.â€ In Handbook of International Historical Microdata, ed. Disseminate anonymized microdata. Note No recommendation regarding prior censuses. IPUMS pays to recover, document, and archive. 5 First Population Census of Sudan 1955/56. Conducted by the British. Time span one year and a half. Used traditional administration (chiefs of the tribes). Total population 10.1 million, adjusted to 10.3 for under-count.Â A population census throughout the Sudan should be conducted and completed by the end of the second year of the interim period. Therefore this census is a constitutional one. 24 The Central Bureau of Statistics (CBS) and the Southern Sudan Commission for Statistics and Evaluation (SSCSE). IPUMS-International: Harmonizing and Disseminating Census Microdata. Ragui Assaaad, University of Minnesota, AUC and ERF. What is IPUMS? Primary Goals of IPUMS. 1. Preserve, archive, and protect 2. Harmonize and integrate 3. Disseminate.Â 11 to 20 Industry Urban-rural status Ownership of dwelling Years of schooling Children ever born Religion Household size Nativity status Number own children in HH Motherâ€™s location in HH. 21 to 30 Spouseâ€™s location in HH Country of birth Fatherâ€™s location in HH Family size Children surviving Number own children <5 in HH Group quarters status Age of eldest child Age of youngest child Total Income. Research Applications. â€¢ Population growth, decrease, movement â€¢ Fertility and Mortality â€¢ Household composition and living.