IPUMS-International 2008 Advisory Board Report

Prepared by the co-investigators and staff of the IPUMS-International project

May 2008

Minnesota Population Center

Contents

Background

In September 2007, we prepared a comprehensive report for the Advisory Board that detailed our progress in seven key areas and described our plans for each area in the coming five years.

(1) creating a broader user community and updating outreach efforts;
(2) acquiring data, and especially securing participation of large countries;
(3) implementing cyber-innovations;
(4) improving documentation of geographic boundaries;
(5) assessing data quality;
(6) processing data; and
(7) developing new data products.

Our plans have not changed greatly since September, so we will not repeat that information here. Instead, we attach the main body of the September report as Appendix A. The present report is therefore limited to a brief summary of our progress from October 2007 to April 2008. We report on the following: 1) a thorough external review of IPUMS-International procedures; 2) our progress in data acquisition and preservation; 3) our latest data release; 4) a major revamping of our user interface; 5) the creation of new "pointer" variables for identifying family interrelationships; 6) our recent successes in data dissemination; and 7) a new initiative related to IPUMS-International—the World Population Data Network.

Trewin Report

In November 2007, Dennis Trewin conducted a comprehensive external review of IPUMS-International data security and confidentiality protection. Trewin is a former head of the Australian Bureau of Statistics and former President of the International Statistical Institute. Prior to the review, he had been skeptical about whether it was possible to disseminate microdata internationally without compromising respondent confidentiality. Trewin conducted a week-long investigation, examining every aspect of IPUMS-International procedures, from the training of technical staff to the physical security of the server room. In the end, Trewin's report was highly favorable, and included the following comments:

- "IPUMS-International is a valuable and trustworthy microdata service. It meets the fundamental principles of good practice with respect to confidentiality and microdata."
- "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."

Trewin had limited recommendations for improvements, and we are taking those recommendations very seriously. The complete Trewin Report is reproduced as Appendix B.

Data Acquisition and Preservation

We have obtained 10 new dissemination agreements since the last report, bringing the total to 80 countries representing 80 percent of the world's population. The new agreements cover high-

density employment and health survey data for India (an agreement on the Indian censuses is still in negotiation). Other Asian governments that have recently signed agreements are Bangladesh, Nepal, Jordan, and Kyrgyzstan. In Africa, we have new agreements with Ethiopia, Tanzania, Zambia, Sierra Leone, and Botswana.

We have now received and preserved microdata for a total of 212 censuses of 72 countries plus 5 large scale surveys of India (See Appendix E). New acquisitions include:

- Asia: India (5 high-density employment surveys), Bangladesh (1991), Nepal (2001), Mongolia (1989)
- Africa: Botswana (2001), Ethiopia (1984, 1994), Mauritius (1990, 2000), Sierra Leone (2004), Tanzania (2002), Zambia (1990, 2000)
- Americas: Haiti (1971, 1982, and defective microdata for 2003), Saint Lucia (1980, 1991)

We are collaborating with Bangladesh and India on two major in-country data recovery operations. In Bangladesh, we will restore approximately 300 9-track tapes; in India, another 200. We are planning smaller recovery operations in Fiji, Nepal, and Jordan, and we anticipate that additional recovery needs may arise. The recovery operation in Mongolia is complete and microdata entrusted for the 1989 census.

We are optimistic about the prospect in the near future of new agreements with several statistical agencies, including Cuba, Hong Kong, Jamaica, Japan, RO-Korea, Morocco, New Zealand, Poland, Switzerland, Timor Leste, and Turkey.

In the coming year, we expect to obtain the first datasets from the 2010 round of censuses. In a meeting of the Principal Investigators with representatives of Latin American statistical agencies in Panama City, scheduled for early June 2008, we expect to collect additional Latin American census datasets, as well as to discuss such matters as the geographic and GIS components of Latin American data integration.

Data Release

In May 2008 we will complete work on 32 additional samples from 14 countries. We are currently putting the finishing touches on the data release, which will both add new countries and extend the data series for countries already in the IPUMS. We are enriching the holdings for existing countries by adding samples for 1990 China, 2005 Colombia, 1995 and 2005 Mexico, 2005 United States, and 2001 Venezuela. In addition, we are replacing the .1% 1982 China sample with a 1% sample. Nine entirely new countries are being added to the data series: Austria, Canada, Egypt, Ghana, Iraq, Malaysia, Netherlands, Panama, and the United Kingdom. This release strengthens our concentration in Europe and creates a new concentration in the Middle East, where four countries are now participating.

Our upcoming data release will add over 3,000 new variables to IPUMS. Of these, 70 are new integrated variables—nearly all of them introduced to handle geographic or ethnic information unique to the new countries. The rest are the unharmonized variables unique to each of the new datasets.

We now have approximately 500 integrated variables and 12,600 unharmonized variables. About 9,000 unharmonized variables are available to researchers. Most of the 3,600 suppressed variables are internal technical variables that we generate; some are undocumented or redundant data; others pose a confidentiality risk, such as highly detailed geographic information. In sum, apart from small-level geography and a handful of highly sensitive ethnicity variables, researchers essentially have access to all the information in the original data files.

As with every data release, we had to revise the documentation—and in some cases the codes—of hundreds of existing integrated variables. To give a sense of scale, the compiled enumeration text associated with the employment status question is now 80,000 words. Marital status is 25,000 words. We had to consider this body of material as we incorporated the new samples and the comparability challenges they introduced. Some variables required harmonizing the coding structures and labels for input variables from over 100 samples.

User Interface

We substantially redesigned the IPUMS web interface this year. In fact, this is the most substantial change in the IPUMS methods for variable display and data extraction since the original IPUMS-USA web system was introduced in March 1996. Several of the new features required redesign of aspects of the data access system, and we used this opportunity to introduce major innovations to the user interface that address future scalability issues.

Although it is not evident to users, we also rationalized the underlying code that governs the website and programmed dozens of unit tests to ensure that the code performs as intended. This reorganization will make maintenance and further development easier as we move forward. Such continuing infrastructural improvement is an essential, if unglamorous, aspect of a large and evolving project.

Variable Browsing. A new menu system gives users greater control over the display of variables. By default, the system now displays one variable group (e.g., "core demographic") at a time. This innovation enables the display of variable availability information across 111 IPUMS samples without bogging down the browser. A prominent menu button toggles between the integrated and unharmonized variables, making the latter far more visible in the new system. Previously, the unharmonized variables were only viewable by sample. We now offer a view of them organized into variable groups—like work or education—in the same way as the integrated variables.

We have introduced an "Options and Help" button that provides a drop-down list with a number of choices. The first item on the list restores the default viewing options for the variables page. The last item brings up extensive help text. Each of the remaining items on the list is a toggle that provides an alternative view from the default behavior. Users can still choose to see all variable groups displayed on a single page, rather than the default single-group view. They can also choose to see summarized availability information (i.e., a count of how many samples include each variable on the list), or to see variables not available in the samples they currently have selected. The options menu provides a convenient place to add more features in the future.

When redesigning the interface, one of our goals was to integrate variable browsing and variable selection for inclusion in extracts. With thousands of variables, it's increasingly inconvenient to identify a relevant variable when outside the extract system, only to have to *rediscover* it when defining a data extract. We therefore added an "Include in Extract" column of check boxes to the variable list. Now users can earmark variables outside the extract system -- not just in the variable list, but also from the variable descriptions and codes pages. Any variable so identified will be pre-selected when the user enters the data extract system. If the user selected samples while browsing the variables, these same samples will be pre-selected for the data extract system as well. Of course, users can change sample selections and unselect variables within the extract process. Selections made outside the extract system do not persist beyond the current web session.

The variables page still lacks a variable search feature, and we consider adding such a search feature to be a high priority. The redesign of the web site became a very large project, however, and we decided to delay this aspect—and a few other lesser improvements—until after the current data release.

Extract Interface. The data extract system has been substantially altered. All of the variable browsing features discussed above are replicated in the extract system, including the ease of switching between integrated and unharmonized variables. The steps in the data extract process have been altered and extended to accommodate new features. In addition, every page of the extract process now has its own help text.

A redesigned first screen serves as the entry and exit point for variable selection when making an extract. This first screen lists the pre-selected variables, allows the de-selection of variables, and serves as a kind of shopping basket to summarize the user's choices before proceeding to later steps of the process.

We added a new, second screen after variable selection to consolidate variable actions in one place. In the old system, case selection was a column in the variable selection list. We moved this function to the second screen, so users could make coordinated choices. The second screen makes the general/detailed variable distinction more evident. Now we offer both the general and detailed versions of a variable to users by default, but they can unselect one (e.g., opting for only the general version of Employment Status) if they wish.

The second screen also provides a place for an important new feature of great value to researchers—a utility for attaching a characteristic of a person's mother, father, spouse, or household head as a new variable on the person's record. For example, using the "Occupation" variable, users can opt to automatically create a new variable for "Occupation of mother." All persons in the extract who reside in a household with their mother would receive a value for this new variable. The extract system automatically generates a name for the new variable. The attached-characteristics feature builds upon the constructed IPUMS "pointer variables" that identify co-resident mothers, fathers, and spouses for each person (discussed further in the following section).

A new, third screen introduces yet another fresh feature of the extract system: the ability to customize sample sizes. In our 2007 Report to the Board, we noted that extremely large samples, and large numbers of cases generated for research projects adopting a cross-temporal or cross-national approach, posed logistical problems for user downloading and analysis. Under the prior system, it was difficult for users to predict the size of their data extracts or to do anything about it. As we promised in our 2007 Report, we have implemented a solution to sample size management—and this solution is present on the third, sample size screen of the extract system.

The sample size screen calculates the expected size of a user's data extract. If this is too large, or if users want to manage the sizes of the individual samples in their extract, a tool allows them to do so. Enter a number of households or persons or a sample density for any sample and the system will draw the proper number of households from that dataset to yield the target value. An "all samples" option applies the same selection to every sample in the extract.

The sampling unit for the sample-size tool is the household. The system will draw a systematic sample of every Nth household at the proper density to produce the number of cases requested. If a user chooses a reduced sample size, the syntax files created by the extract system contain programming to alter the household and person weights to reflect the new sample densities. A link on the sample selection screen provides extensive help text to assist users in judging whether and how to alter sample size from the default full sample size option.

Family Interrelationship Variables

In 2006-2007 we developed the location-of-spouse variable that identifies the person number within the household of each married person's co-resident spouse or partner. This year we constructed the location-of-mother and location-of-father variables. This work was much more difficult than the spouse pointer. It required a great deal of innovation to handle the design, logistics and testing of algorithms for the nearly 100 samples that had the requisite input data. Roughly half of the person-hours spent on the project in the past year went into this work. Fortunately, we now have a method and design that will allow us to add this feature to additional samples much more cheaply in the future.

Using 200,000 person subsamples of every census, we developed our core algorithm, which differs considerably from the one we developed for a limited set of international samples in 2003-2004. There is too much variation in the reporting order of the enumerated persons, the categories of the relationship-to-head information, the marital status information (sometimes including consensual unions and polygamy), and the quality of the data for our experience with IPUMS-USA to offer more than rough guidance. Despite the challenges, we decided to create links for all persons, not just those under age 19 as we had in our prototype parental pointer links.

We applied a series of rules that looked for particular relationship-to-head pairings, starting with the simplest and ending with the most ambiguous relationships. Each person is evaluated individually, searching through the household for a potential mother and father. We integrated the mother and father links, but at any given "strength" of test, the mother is sought first. When a mother or father is linked to a child, the mother's or father's spouse is also linked at that time.

Methodologically, we first explored the data and developed a basic set of rules that minimized obviously bad links while maximizing links for young children. After that, we set up a "voting" system in which every proposed edit to the algorithm was independently evaluated by the staff, with each person deciding if any link that changed was better, worse, or indeterminate from the previous version of the algorithm. We tested 26 proposed changes in this way, and compared the final result to what we had at the start.

We made a number of notable decisions in developing the parental pointers. We used, but de-emphasized, the order of persons within households as a predictor of relationship. Some samples with complex household structures lacked fertility variables, and we never have such information for men. To keep one potential mother or father from monopolizing all children in such situations, we developed a synthetic "cap" based on the number of children and available parents. Ever-married parents take precedence, but they will not necessarily take all the children. The allowable relationship pairings had to be customized in many instances. One such challenging situation was when a relationship like grandchild was grouped with "other relatives." In other situations we had to develop rules to decide which of an undifferentiated "child/child-in-law" couple was the blood relative of the head. The children of unmarried partners and polygamous households provided further challenges.

There are limitations to evaluating parental links on a person-by-person basis without some kind of pre-pass to characterize the household. But, after much consideration, we determined that a system using some kind of artificial-intelligence model could not be developed with the time and resources available. Such a system is quite possibly where we will need to go in the future, to make the next evolutionary step.

Dissemination and Publications

One of the mandates we received from the Advisory Committee was to create a broader user community. In our prior report of September 2006, we extensively documented our efforts to meet this goal. Here, we can further report that the number of approved users of the database grew by 55 percent in 2007 and now exceeds 2,000. IPUMS-International is thus one of the most widely-used restricted access data collections in the world.

The robust growth in the number of database users can in part be attributed to our active program of training and dissemination. We not only offered our usual summer training workshop at the University of Minnesota but also extended our training program internationally. In December 2007, we conducted a multi-day, hands-on data users workshop in Arusha, Tanzania, in conjunction with the meeting of the Union of African Population Studies. We also held a data producers workshop at Columbia University, in which we discussed data integration and dissemination plans for the 2010 round of censuses (see Appendix F for the list of attendees).

We exhibited information about IPUMS-International at the American Sociological Association, the International Statistical Institute, the Allied Social Sciences Association, the Population Association of America, and the American Public Health Association. Presentations about the IPUMS-International project and substantive papers and posters based on these data for scholarly conferences by members of the staff also brought word of the database to a larger audience. (See the bibliography in Appendix C for more details.)

In May, 2008, we invited registered users of IPUMS-International data to submit papers for the newly-established IPUMS-International Research Awards. The two awards—for the best published paper and the best paper by a graduate student—were established to showcase the strength of the IPUMS-International database for cross-temporal and cross-national research. Each award consists of a cash prize of $250 and plaque, and winners will be invited to the Minnesota Population Center to present their research at a special symposium. Information about the awards was not only sent to all registered IPUMS-International users but also is prominently displayed as a news item on the portal page (www.ipums.org) for MPC datasets generally.

Our data dissemination efforts are beginning to bear fruit. The number of journal articles, books, dissertations, and working papers based on IPUMS-International increased 89% in the six months since the last report. The current bibliography, with nearly 200 items, appears in Appendix C. Scholarly works using IPUMS-International data have now appeared in such leading journals as *Demography* (2 articles), *Population and Development Review* (3 articles), *Social Biology*, *International Migration Review*, and *Labour Economics*. Publications based on the database are also reaching a broad international audience, through both international conference papers and through such journals as *Les Cahiers Quebecquois de Demographie*, *Latin American Research Review*, *Eastern European Economics*, *European Journal of Population*, *Revista Brasileira de Estudos de Populacao*, *Scandinavian Population Studies*, and *Romanian Journal for Population Studies.*

World Population Data Network

In October 2007, the National Science Foundation Office of Cyberinfrastructure announced a new $100 million program entitled "Sustainable Digital Data Preservation and Access Network Partners (DataNet)." The central goals of the initiative were to

- provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline;
- continuously anticipate and adapt to changes in technologies and in user needs and expectations;
- engage at the frontiers of computer and information science and cyberinfrastructure with research and development to drive the leading edge forward; and
- serve as component elements of an interoperable data preservation and access network.

Since these goals accord closely with our own, we quickly moved to assemble a team from across the country and around the world that could effectively respond to the call. Our project, the World Population Data Network (WPDN), involved a wide variety of data sources, but the IPUMS-International collection provided the unifying theme and rationale.

We submitted a preliminary proposal in January 2008 and were among seven institutions invited to submit a full proposal in March 2008. We were ranked fourth of those seven proposals, and only the top three received site visits. Sylvia Spengler, the program officer, strongly encouraged us to resubmit for the second round of the competition, scheduled for October. We attach the WPDN proposal as Appendix D.

Planning for 2009-2014

Our September 2007 report (Appendix A) outlined our plans for the next IPUMS-International funding period. We need to refine our priorities, since there is too much to get done in a five-year period. Our objectives include:

- ***Preservation.*** We have made great strides in our efforts to preserve the world's statistical heritage, but the job is not complete. We believe this should remain a top priority for the next five-year funding period.

- ***Data production.*** By 2009, after ten years of effort, we expect to have released over 130 census samples. We anticipate, however, that the scale of the collection available for dissemination will grow dramatically. We expect to have the raw materials available for well over 300 samples within the next five years, and we should make as much of it available to scholars as possible.

- ***Redesign of variables, metadata, and work processes.*** At the outset of the current grant in 2004-2005, we completely redesigned the data collection, metadata structure, and work processes for data production. This kind of redesign must be undertaken on a regular basis; to accommodate the expanded scale of the collection, we will have to do it again.

- ***Data quality assessments.*** As described, we plan to undertake a systematic program of data quality assessment.

- ***Cyberinfrastucture.*** The WPDN described a much more expansive vision for cyberinfrastructure than we included in our September 2007 report. We believe these ideas would yield major benefits, particularly as the scale of the database continues to expand. If WPDN is not funded, we will have to make hard decisions about investment in cyberinfrastructure for IPUMS-International.

- ***Geographic improvements.*** The geographic improvements we have described are essential to realize the full potential of the data collection, but they will be expensive.

- ***New data products.*** The complete-count microdata with full geographic identifiers and the internationally-compatible detailed summary files for small areas are both new kinds of data that have never before been available. They both have great potential to address questions of vital importance.

We solicit the input of the Advisory Board as we prioritize these and other needs for the next five-year period.

Appendix A
IPUMS-International Response to Advisory Board

September 2007

Background

This document responds to questions posed by the IPUMS-International Advisory Board following our 2007 meeting. We report on our progress over the past year, our plans for the final year of the current grant, and the directions we would like to take during the five year period from 2009 to 2014.

The IPUMS-International project began in 1999 with a social science infrastructure grant from the National Science Foundation (SBR-9908380). Our goal was to demonstrate the feasibility of preserving and harmonizing census microdata from around the world and making them accessible to researchers. That grant resulted in the integration of 28 census samples from eight countries, using methods largely adapted from the earlier United States IPUMS project (SES-9118299).

In 2004, we received funding to greatly expand the scope of the data series. In addition to grants from the National Institutes of Health, we were awarded a major infrastructure grant from the NSF Human and Social Dynamics priority area (SES-0433654). The National Science Foundation appointed an Advisory Board to oversee the project, chaired by Tim Smeeding (Syracuse University) with Doug Anderton (University of Massachusetts, Amherst), Gyimah-Brempong Kwabena (University of South Florida), Bill Lavely (Jackson School of International Studies, University of Washington), and John Logan (Brown University). The National Science Foundation was represented on the Board by Dan Newlon (Economics) and Patricia White (Sociology).

In December 2004, the investigators met with the Advisory Board and outlined a plan to revamp the work process, metadata, and software infrastructure of the project. This redesign was necessary, since the handcrafted methods used to produce 28 samples under the first NSF IPUMS-International grant were unsuited to a project promising 100 additional samples in a similar five-year period. The web site also required substantial revision to make the increased volume of material manageable for researchers.

We spent most of 2005 redesigning the technical infrastructure of the project. We broke sample processing into a series of discrete steps amenable to the employment of a much expanded staff. Key to the overall redesign is a metadata-centric approach, in which the research staff manipulates relatively simple but highly-structured documents that drive the data processing and 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 framework of our redesigned metadata systems and work processes was in place when we met with the Advisory Board for the second time in May 2006. At that meeting, we described the

new method of data processing and some of the metadata tools we had developed, and we demonstrated a new web interface that compiles variable-specific enumeration materials dynamically on the web. At the time, the new features existed only on our development web site, and we were still in the midst of preparing our first data release under the revised regime.

At the 2006 meeting, the Advisory Board made specific recommendations concerning the priorities of the project. These included:

- Release more data. After the investment in project infrastructure, it was time to produce many samples.
- Make unharmonized variables publicly available. The proper balance of effort between harmonized and unharmonized variables is difficult to determine in the abstract, but there was consensus on the desirability of making the source information available.
- Provide more geography. Users should be given as much geographic detail as the subject countries will allow.
- Increase marketing and outreach to attract more users. More information should also be collected on the users, to inform decision-making.

By our third Advisory Board meeting in May 2007, we had addressed the Board's recommendations. Most important, we more than doubled the number of IPUMS-International samples and increased the number of variables more than ten-fold, demonstrating the dramatic productivity improvements made possible by our new software infrastructure and redesigned work process. In addition, we further streamlined our processing procedures and software, added important new website features, greatly improved the level of geographic detail in the samples, and took important steps to improve marketing and outreach.

Because of this substantial progress, the Advisory Board recommended that the National Science Foundation cancel a site review that had been planned for the summer of 2007. Instead, the Board requested that we prepare this written report on our progress and plans both for the final two funded project years and for the longer-run future. The Board specifically requested that we address the following topics:

(1) creating a broader user community and updating outreach efforts;

(2) acquiring data, and especially securing participation of large countries;

(3) developing cyber-innovations;

(4) improving documentation of geographic boundaries; and

(5) assessing data quality.

The discussion that follows addresses each of these areas in turn. For each topic, we describe our recent progress and our plans for the final two project years. Where applicable, we also describe our long-run goals for the 2009-2014 period. We conclude with two additional sections not specifically requested by the committee—data processing and new data products—because we anticipate that these areas will be major components of a continuation proposal.

1. Dissemination and outreach

The long-term NSF investment in IPUMS-International is only justified if the data are widely used to produce important new discoveries; accordingly, investment in dissemination and outreach is essential. Beginning with our first data release, the project has actively pursued several strategies to inform the research community about the project. We initially publicized the database by an email announcement to the large user list for the more established IPUMS-USA database. We also announced the international data on the high-traffic IPUMS-USA website and other related websites. Since September 2004, the Minnesota Population Center (MPC) has provided exhibits featuring IPUMS-International at 24 major conferences around the world (see Table 1). We have found these exhibits invaluable not only to introduce

Table 1. Conference exhibits featuring IPUMS-International: 9/2004-8/2007

2004	
	Asociación Latinoamericana de Población, Caxambú, Brazil, September 1
	Social Science History Association, Chicago, November 18-21.
2005	
	Joint Statistical Meetings, Minneapolis, August 7-11.
	International Union for the Scientific Study of Population, Tours, France, July 18-23.
	International Congress of Historical Sciences, Sydney, Australia, July 3-9.
	Seminario Internacional de Población y Sociedad, Salta, Argentina, June 8-10.
	International Statistical Institute, Sydney, Australia, April 5-12.
	Population Association of America, Philadelphia, March 31-April 2.
	Association of National Census and Statistics Directors of America, Asia, and the Pacific, Seattle, March 7-9.
	American Economic Association, Philadelphia, January 7-9.
	Social Science History Association, Portland, November 3-6
	American Sociological Association, Philadelphia, August 13-16.
2006	
	Social Science History Association, Minneapolis, November 2-5
	American Sociological Association, Montreal, August 10-14
	Society of Labor Economists, Cambridge, MA, May 5-6
	Population Association of America, Los Angeles, March 30-April 1
	European Social Science History Association, Amsterdam, March 22-25
	European Population Conference, Liverpool, England, June 2006
2007	
	International Statistical Institute, Lisbon, August 21-30
	American Sociological Association, August 11-14
	Organization of American Historians, Minneapolis, March 29-April 1
	Population Association of America, New York, March 29-31
	Allied Social Sciences Association in Chicago, January 5-7
	American Historical Association in Atlanta, January 4-7

the database to new users, but also to establish face-to-face contact with our existing users and to obtain feedback from them. At most of these conferences, we have also participated by presenting papers that describe aspects of IPUMS-International or use IPUMS-International data.

We have conducted a variety of training workshops to help users fully exploit the power of these large-scale datasets. In July 2006, January 2007, and July 2007, we held multi-day IPUMS workshops in Minneapolis. The workshops, which covered both IPUMS-International and IPUMS-USA, combined presentations with hands-on laboratory work. Topics included sample designs, database creation, and data extract systems, weights, geographic variables, measurement of socioeconomic status, and constructed variables. Each workshop could accommodate 30 people but elicited more than 100 applications. This demand is extraordinary, considering that our sole publicity was a single e-mail to our active users, and that participants had to pay tuition, travel, and local expenses in Minneapolis. IPUMS-International has also been covered each year at two-hour data workshops held in conjunction with the American Sociological Association. In 2006 and 2007, we took advantage of two national conferences held in Minneapolis to conduct special half-day training workshops with minimal travel costs. With financial support from the Norwegian government, we participated in a training workshop in Tanzania in January 2007. A staff member presented the IPUMS data series to a team of Tanzanian researchers at the University of Dar es Salaam. The two-day workshop covered the design and scope of the data series, and it included a data analysis component. The participants were highly enthusiastic, and several of them have become active users.

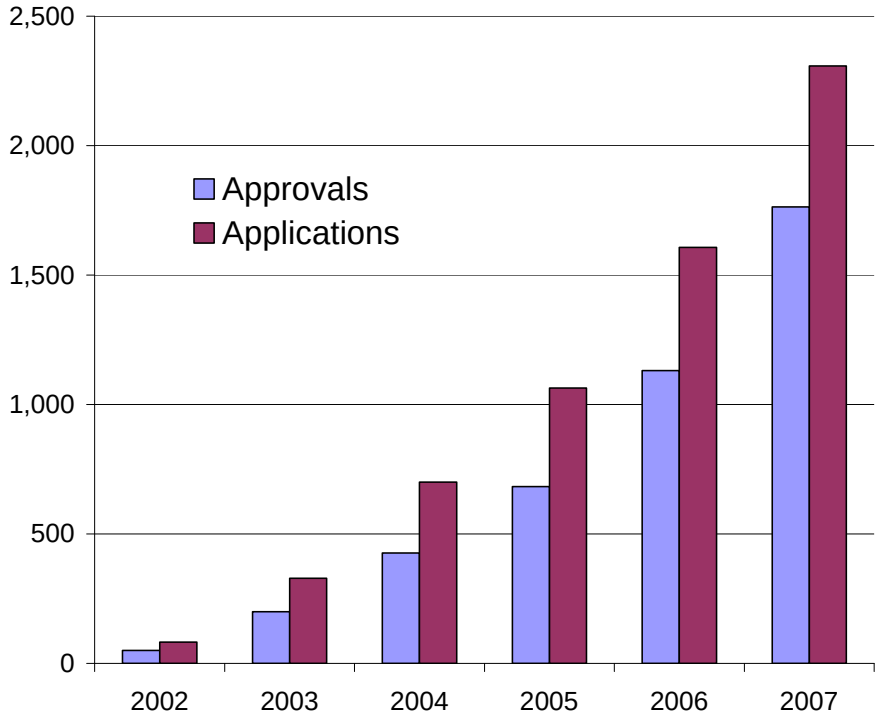
We have enlisted the assistance of data producers to reach new users. We have held multi-day data producer workshops in Seattle (March 2005), Paris (June 2006), and Lisbon (August 2007). While the primary purpose of these workshops is to communicate with our partners and obtain new data and data dissemination licenses, we also used these opportunities to develop dissemination strategies in concert with national statistical agencies and other international partners. For example, at the Paris workshop, we discussed plans for a specialized website designed for European users of the database. That website is now online (http://www.iecm-project.org/).

These dissemination efforts have been extremely successful. Figure 1 shows the number of applicants and approved users in each year since the first data release in 2002. About a quarter of applications for use are rejected because the proposals do not meet the requirements of our dissemination agreements with national statistical agencies.

For the past three years, the number of users has been growing at a pace of approximately 60 percent per year. This is a substantially faster pace of growth than we saw with IPUMS-USA data a decade ago. It is reasonable to infer that, with continued support, the database will become one of the most widely-used data sources in the social sciences.

To put the IPUMS-International usage in perspective, we compared it to other large datasets of similar vintage. Table 2 shows the number of users as of July 2006 for several multi-million dollar data collection projects that released their first data in 2002 or 2003. These statistics come from a survey of data producers conducted a year ago by the Demographic and Behavioral Sciences Branch of NICHD. For the most part, these projects cost several times as much as

Figure 1. Number of IPUMS-International applicants and users, 2002-2007



IPUMS-International, but they all had significantly lower early usage. One reason, we expect, that these data collections have attracted fewer users is that the other projects produced comparatively small and topically specialized samples. We believe, however, that our aggressive program of outreach and training has also played an important role in the early success of our data dissemination program.

Table 2. Number of users as of July 2006 for recent high-cost data projects

Dataset	Release	Users
Fragile Families	Jun-02	375
New Immigrant Survey	Jun-02	335
IPUMS-International	Aug-02	915
Three-City Study	Aug-02	33
LA-FANS	May-03	611

We hope to expand our dissemination and outreach programs in coming years. In August 2007, we submitted an R25 training grant proposal to the National Institutes of Health. If successful, this grant will provide training on the use of data from IPUMS-International and IPUMS-USA to approximately 950 researchers over the next five years. The centerpiece of the program is an intensive five-day workshop offered once each year to 30 participants. These workshops will be tailored to highly promising early-career scholars who already have a solid background in

statistics and data analysis. We will target a diverse group of sophisticated users with the greatest potential to produce path-breaking interdisciplinary demographic research. In addition to mastering the use of the databases, participants will work on individual research projects, network with others who have similar research interests, and hear from exemplary data users who will serve as mentors and models for using these data creatively and effectively.

The proposed R25 training program also includes shorter introductory workshops that will reach a larger audience. These include one-day and half-day sessions at conferences and on-site at the Minnesota Population Center, to focus on issues commonly arising for beginning users of frequently-requested IPUMS samples. These short courses will improve the efficacy of new users, increase the diversity of the user community, and introduce junior researchers to online resources that will provide further learning opportunities.

The workshops will also provide a venue for developing and refining introductory and advanced educational materials on using the datasets properly and efficiently. These materials (including tutorials, user notes, and practice exercises) will be made available on the web, and thus serve, in a cost-effective manner, a large and growing community of researchers using IPUMS-International datasets.

Beyond the web-based instructional material, we plan additional web-based features that will capitalize on the extensive knowledge base of our users to create user communities and assist with user support. We also plan to develop new dissemination software—such as online tabulation—that promises to expand the audience for IPUMS-International data. These innovations are described below in section 3, "Cyberinfrastructure."

2. Data Acquisition

The Advisory Board inquired about our plans for acquiring new partners, and especially for securing participation of large countries. We have had remarkable success in the past year in obtaining new data and dissemination agreements. At the time of our 2006 progress report, we had data from 140 censuses in 46 countries. At this writing, we have received 185 samples from 63 countries, and we have signed agreements with 70 countries. The 70 countries that are now participating in IPUMS-International have a combined total population of over four billion, or about 61 percent of the world total. We have already exceeded the data acquisition we described in our 2004 IPUMS-International HSD grant proposal by a margin of 20 percent. Nevertheless, we consider it vital that we continue our efforts. Indeed, we believe that preserving and opening access to these irreplaceable data resources is the most important contribution of the project. If our current negotiations are successful, we will obtain about 280 censuses from 100 countries, not counting future censuses from the 2010 round.

Our data acquisition priorities are based on a number of criteria. For example, we prioritize countries whose data are at high risk of destruction, that have a long run of surviving high-quality censuses, or that have undergone dramatic demographic or economic changes during the period covered by their surviving censuses. From the outset of the project, however, we have also devoted special attention to acquiring data and dissemination agreements from large countries. The cost of acquiring and processing a census is virtually the same regardless of the

size of the sample, so focusing on the most populous countries is cost-effective. Moreover, data from the largest countries often generate the greatest excitement in the research community.

Our efforts to obtain data from populous countries have been notably successful. Table 1 shows the participation of the world's 30 most populous countries in the world. Countries for which we have signed dissemination agreements are shown in bold, and those currently under negotiation are in italics. We have already received data from four of the five largest countries. Much of our success with the largest countries has occurred in the past six months. For the largest country, China, we recently acquired two new samples that boost the number of cases twenty-fold. Data for Indonesia, Pakistan, and Bangladesh are also now in hand. Therefore, among the seven largest countries, there is now only one non-participant.

The missing country, however, is an extremely important one; India will soon be the largest country in the world. High-quality Indian microdata have survived for 1991 and 2001, and perhaps for 1981. Acquiring Indian data is therefore our highest priority. Robert McCaa visited India in the summer of 2007, and plans an extended visit in winter 2007/8. Indian census microdata have never been made available outside the Census Commissioner's office. There will soon be a new Census Commissioner, and we are hopeful that new leadership may offer new opportunities for cooperation. A workshop in New Delhi of the census chiefs of the South Asian Regional Commission (SARC) to be held in January or February 2008 will offer an ideal opportunity to display our excellent relations with the National Statistical Offices of Bangladesh, Pakistan and Nepal as well as inform the new commissioner of the benefits of the IPUMS project.

As shown in Table 2, we have signed dissemination agreements with the national statistical offices of 21 of the largest 30 counties, and we have the data in hand for all but one of these countries. Of the remaining nine countries, we are actively negotiating with five, with high hopes of eventually securing agreements. Four countries among the top 30 are presently on hold: Japan, Iran, Myanmar, and Ukraine. Japan and Ukraine have rejected our proposal, but we plan to approach both countries again in due course. We will approach Iran and Myanmar as the political situation permits.

Regional meetings of national statistical offices, like the one we are planning for New Delhi, have been key to our success in acquiring census microdata. Such meetings offer valuable forums to showcase our accomplishments, strengthen existing partnerships, and invite participation by national statistical offices not yet affiliated with the IPUMS initiative. We are now regularly invited to make presentations at regional census directors' meetings in Latin America, the Caribbean, Europe, Asia, Africa, and now, the Arab States (Amman, Jordan, Nov 12-13, 2007).

Beginning in 2008, the IPUMS project will be represented by an official delegate on the floor of the annual meetings of the United Nations Statistical Commission in New York City. In addition to networking, the delegate will be able participate in debates on issues of critical importance to our continued success. IPUMS is the first academic project to garner the accolade of "good practice" from the UN-ECE (2007). Fortunately, as the premier provider of census microdata world-wide, we are well positioned to influence the debate, as the UNSC completes its recommendations on managing access to census microdata.

`Table 4. Status of 30 largest countries

Rank	Country	Population	Status
1	**China**	**1,321,851,888**	**Disseminating**
2	*India*	*1,129,866,154*	*Negotiating*
3	**United States**	**301,139,947**	**Disseminating**
4	**Indonesia**	**234,693,997**	**Data Received**
5	**Brazil**	**190,010,647**	**Disseminating**
6	**Pakistan**	**164,741,924**	**Data Received**
7	**Bangladesh**	**150,448,339**	**Data Received**
8	*Russia*	*141,377,752*	*Negotiating*
9	*Nigeria*	*135,031,164*	*Negotiating*
10	Japan	127,433,494	Inactive
11	**Mexico**	**108,700,891**	**Disseminating**
12	**Philippines**	**91,077,287**	**Disseminating**
13	**Vietnam**	**85,262,356**	**Disseminating**
14	**Germany**	**82,400,996**	**Data Received**
15	**Egypt**	**80,335,036**	**Processing**
16	**Ethiopia**	**76,511,887**	**Data Received**
17	**Turkey**	**71,158,647**	**Agreement signed**
18	*Congo*	*65,751,512*	*Negotiating*
19	Iran	65,397,521	Inactive
20	**Thailand**	**65,068,149**	**Data Received**
21	**France**	**63,718,187**	**Disseminating**
22	**United Kingdom**	**60,776,238**	**Processing**
23	**Italy**	**58,147,733**	**Data Received**
24	*Korea, South*	*49,044,790*	*Negotiating*
25	Myanmar	47,373,958	Inactive
26	Ukraine	46,299,862	Inactive
27	**Colombia**	**44,379,598**	**Disseminating**
28	**South Africa**	**43,997,828**	**Disseminating**
29	**Spain**	**40,448,191**	**Disseminating**
30	**Argentina**	**40,301,927**	**Disseminating**

In addition to negotiating agreements with additional countries to distribute microdata, we are constantly negotiating with current participants to obtain additional datasets. In some cases, these additional data are historical samples that need processing or recovery before they can be released. In other cases, we seek larger or more detailed samples of the censuses we are already distributing. In virtually every country, we must negotiate to obtain data from the 2010 round of censuses. This acquisition work will continue for at least another five-year funding period.

We do not have the resources in the current grant to fully process the flood of microdata arriving in Minnesota, but we are taking basic steps to ensure preservation. For each census, we make sure that the files are readable and complete and that the documentation corresponds to the files

received. We then encrypt the data and store it securely on-site and off-site, taking precautions to ensure protection from both disclosure and data loss.

3. Cyberinfrastructure

Overview of existing IPUMS-International cyberinfrastructure

This project has required us to develop a substantial body of new software and metadata.[1] IPUMS-International software can be grouped into four principal categories:

- ***Metadata preparation software*** is a library of utilities that allow research staff to create and maintain the XML structured metadata that describe every aspect of both our source data and the IPUMS-format data we disseminate. We developed most of this software in 2005 and early 2006, but it is continuously refined and improved.

- ***Data preparation software*** is a set of programs for pre-processing IPUMS-International datasets. These programs are used to reformat samples from their native structure into a consistent hierarchical column format; carry out data integrity checks; implement logical edits to correct structural errors in the data; draw samples; perform dwelling-level substitution to eliminate unusable cases; and impose confidentiality measures.

- ***Data conversion software*** is a system that recodes the pre-processed data into IPUMS format; creates a range of standard constructed variables including the IPUMS family interrelationship pointer variables; carries out variable-level logical edits; allocates missing or inconsistent data items; and generates frequencies for each variable. We revised this software substantially in 2005 to operate on a new XML-based metadata structure. We have also added a procedure to identify all differences in the output files produced between successive runs on the same dataset. This allows us to confirm quickly and easily that corrections to the data are successful and that no new errors are introduced.

- ***Dissemination software*** is a suite of programs that provide integrated web access to all data and documentation, allowing users to merge datasets, select variables, and define population subsets in an information-rich environment. The system also allows users to revise previous extract requests and modify old extract specifications to formulate new queries. The web system is password-protected, limiting access to approved users per our international contractual obligations. Improvements under development will offer advanced tools for navigating documentation, defining datasets, and constructing customized variables. In 2005, we replaced the PHP script initially used for IPUMS-International dissemination with a new Java-based system. Like the data conversion program, the new dissemination system operates on a new XML-based metadata structure. In addition, we replaced hundreds of pages of static HTML pages with dynamic documentation pages generated on the fly.

[1] Design of these systems was carried out under the direction of Peter Clark, Monty Hindman, Catherine Ruggles, and Matt Sobek; the software engineers were Marcus Peterson and Colin Davis. The design benefited greatly from the input of Jaideep Srivastava of the University of Minnesota Department of Computer Science and Jeffrey Naughton of the University of Wisconsin's Department of Computer Science, as well as Nupur Bhatnagar, a Minnesota Computer Science graduate student.

All the software for data preparation, data conversion, and dissemination is driven by metadata. Metadata is formally structured documentation of digital data. We have developed a comprehensive metadata system for IPUMS-International, with a goal of capturing everything we know about the data in a structured format that can be processed by machine. Our specification is in some respects similar to the Data Documentation Initiative (DDI) Document Type Definition developed by a consortium of data archives and producers, but it handles additional kinds of metadata required by our project.[2] The IPUMS-International metadata format is compatible with DDI, and we can generate DDI-compliant codebooks for datasets on demand.

Like the DDI, our metadata specification is written in the eXtensible Markup Language (XML). The metadata has a structured format in which each piece of information is identified by a tag that identifies the particular kind of information. For example, there is a tag to indicate that a particular string represents a value label, and another tag to identify the variable universe.

The metadata specification has five major components:

- ***Source data dictionaries.*** For each source dataset, this metadata component provides variable labels and value labels in both the original language and in English, along with input column locations, variable widths and formats, and frequency distributions.

- ***Variable translation tables.*** This metadata component provides most of the variable-level information required to create the database, including IPUMS-format variable labels, value labels, and codes, as well as dataset-specific information on universe, location of source variable, and all information required to harmonize codes across datasets.

- ***Variable descriptions.*** This component provides information for users about each variable and its comparability across datasets.

- ***Control files.*** This metadata component provides information needed to operate and control both the data conversion program and the web dissemination system. Five different control tables identify the symbolic location of each piece of data, metadata, and software needed by the system and control numerous options for the creation and display of each dataset and variable.

- ***Ancillary documentation.*** This component provides information on enumeration instructions and forms in the original language and in English translation, sample designs, and other material related to the particular census or sample.

Software improvements, 2006-2007

Unharmonized variables. The IPUMS software systems now distinguish two classes of IPUMS-International variables. *Integrated* variables are coded in a compatible format across time and space, and are accompanied by extensive documentation of comparability issues. *Unharmonized*

[2] The DDI is described at http://www.icpsr.umich.edu/DDI/.

variables are specific to each census sample, and are coded approximately the same way in IPUMS as they were in the original source.

The system now provides access to over 5000 unique sample-specific variables for public browsing and data extraction, representing virtually all the information in the original samples. Some variables are still suppressed because of obvious data errors or for confidentiality reasons. This represents a milestone in scientifically sound practice, since researchers can in most instances reengineer our data manipulations. Access to unharmonized variables also serves as a practical safety net for the project; if we have misinterpreted something during our harmonization efforts, users can now work around the shortcoming.

Web dissemination tools. The expansion of IPUMS-International necessitated redesign of the data access system. All variable documentation is now generated dynamically and can be filtered based on user-defined selections of samples. When users go to the main variable availability page, they can select all or any combination of samples to display on the screen. The selections persist through the rest of their session. As they browse the system, only the portion of the variable discussions or codes pages applicable to those samples appears on the screen. Users can change those selections at any point, enabling them to control the level of information. The system also compiles relevant enumeration materials for any variable, restricting the output to the samples defined by the user. By filtering out irrelevant documentation and presenting only material for the years and countries of interest, the system allows users to efficiently navigate and comprehend the metadata.

General and detailed variables. For complicated variables, it is impossible to construct a single uniform classification without losing information. Some censuses provide more detail than others, so the lowest common denominator of all samples inevitably loses important information. In these cases, we construct composite coding schemes. The first one or two digits of the code provide information available across all samples. The next one or two digits provide additional information available in a broad subset of samples.

Many integrated variables in the IPUMS have complex coding schemes, due to the variety of classifications in the source data. It takes three digits, for example, to encompass the range of permutations of marital status into a logically organized hierarchical coding structure, but this level of detail is not needed for most analyses.

Accordingly, in December 2006, we added a feature to the data access software that distinguishes general and detailed versions of variables. All integrated variables have a fully detailed version; many now also have a "general" version that utilizes only the first one or two digits of the variable. For example, researchers can access an internationally comparable 1-digit general version of "employment status," or they can use the fully detailed 3-digit version if their research requires finer distinctions. The two sets of codes are completely consistent with one another; one simply provides more categories, while the other is simpler to use and more comparable across samples. Both general and detailed versions of a variable can be included in a data extract.

New registration system and user database. We developed new software to handle applications for use of IPUMS-International data and maintain records on users. These changes were needed

to reduce disclosure risk, improve our capability to analyze data usage patterns, and allow us to implement new data access features. Registrations now expire after one year but are renewable. The approval process for new applications is now handled by computer, which speeds approval time and provides an electronic trail to safeguard against error. The database will record all interactions with users, providing an invaluable resource for optimizing metadata and data access tools. We also now have the capability to record user preferences that persist between visits to the site, although that feature is not yet implemented (see below).

Improvements planned for 2008-2009

On-line tabulation. We will implement on-line data analysis capability using the Survey Documentation and Analysis (SDA) system developed at Berkeley. The system has more than enough analytical capability, but there is considerable work involved in developing a web interface suited to our microdata. The on-line tabulation utility will substantially broaden access to the database.

Advanced extract features. We will make it possible for researchers to construct a variety of variables that capitalize on the hierarchical structure of the data, by expanding the flexibility and functions of the data extraction system. Among the capacities to be developed are the following:

- A procedure for attaching characteristics of co-resident persons (e.g., household heads, family heads, spouses, own mothers, and own fathers) to each individual's record. For example, the system will allow analysts of marriage to create new variables within the extraction system that describe ego's spouse's age or birthplace.

- A procedure for counting the number of persons within each household, family, or own-child group of each parent who have a specific combination of characteristics. For example, the system could count the number of teenage daughters in the labor force for each mother with co-resident children.

Sample size management. Some of the samples in the database are extremely large. Their absolute size can pose logistical problems for user downloading and analysis, and their size relative to other samples poses other practical difficulties in multi-sample extracts. There are at least three strategies for dealing with this: 1) construct equal size subsamples for every dataset in IPUMS-International, and offer that set of samples as an option in the extract system; 2) make smaller subsamples of only the large samples, such as a 1 percent subsample of the 10 percent Mexico 2000 dataset; 3) develop the capacity to have the extract system pull out a subsample on the fly of any specified density, or of a target sample size. Each approach has different benefits and development costs. We will explore the options and implement at least one of these solutions in the current development period.

User preferences and dynamic content. The current system allows users to filter the content of most web pages to contain information relevant only to the samples of interest to them and to set a variety of other preferences. The existing system does not, however, remember preferences from one session to the next. As noted above, have already added the capability to record user preferences to the user database; now we mist add it to the data access system. From any point in the web site, users will be able to modify their preferences for the current session or for a persistent set of preferences, until they choose to edit their choices again.

Dataset IDs for extract replication. The codebook file for each data extract created by our system will include a unique ID as part of the suggested citation. Registered users will be able to enter the ID and draw an identical extract from the data system, thus enhancing the ability of scholars to replicate the results of other researchers.

Integrated variable browsing and selection. Although the interfaces for browsing variables and for performing data extraction are both efficient, they are not interconnected. One can identify variables of interest while browsing but, short of writing down the variable names, there is no way of marking those variables for data extraction. We need to allow users to drop variables into a list, much like a shopping basket. Then, when users move into the data extraction phase, those variables are already pre-selected. At that time, users can choose to drop or retain the pre-selected variables.

Missing data allocation. Many of the IPUMS-International samples—especially those from developing countries—incorporate no procedures to account for item nonresponse or inconsistencies among variables. We will use both probabilistic imputation and logical editing of records to improve the reliability of estimates derived from these census samples. We will focus on those variables that are most used by researchers and most likely to generate logical inconsistencies. When we allocate or edit the data, we will indicate the altered records with appropriate data quality flags, which will be made available through the data extraction system.

Cyberinfrastructure needs for 2009-2014

Even though we have completely revamped IPUMS-International software and metadata infrastructure over the past 24 months, we recognize that this is a temporary solution. IPUMS-International is already the largest collection of accessible population microdata in the world. By 2014, we believe it will comprise a billion records from 250 censuses, and will serve tens of thousands of researchers. To accommodate this massive expansion and improve the efficiency of data production and delivery, we must continue to innovate.

Some of the needed innovation will be improved technical infrastructure. As the quantity of metadata expands, for example, we will need to develop new methods for metadata storage and retrieval, since our current approach will not scale. We will also need to improve the capacity and speed of our systems for data extraction. Professor Jaideep Srivastava and several graduate students in Computer Science at the University of Minnesota are already working on preliminary redesigns for the metadata and data infrastructure underlying IPUMS-International.

We also need to innovate in the development of data sharing technology. Data sharing is central to the project; effective data distribution is essential if the data are to be widely used. The original IPUMS project pioneered web-based data access for large-scale datasets in 1995, and our integrated web-based system for disseminating data and documentation has served as a model for many other social science data projects. As previously described, we have now replaced that aging data access software with a new Java system driven by XML metadata. The current system allows users to merge datasets, select variables, and define population subsets in a rich informational environment. It also offers a robust platform for building new capabilities—outlined below—that will save researchers' time, reduce errors, make replicating studies easier, and democratize access to the database.

1. Social Web. Because of the explosive growth of IPUMS-International research, requests for user support have expanded rapidly. To meet the growing demand, we propose to use new web technologies that leverage the expertise of the IPUMS-International research community. There are currently almost 2,000 users, many of whom are enthusiastic and have great expertise in data from particular countries. If current trends continue, the database could easily have 20,000 users by 2014. We will develop tools and systems that allow users to support each other and improve our website, so less individualized user support is needed. Our goal is to go well beyond the limits of conventional user support, and to foster research communities that focus on specific substantive areas, such as migration, labor force participation, health, or use of data in the classroom. By promoting interaction among users researching similar topics, these communities will provide intellectual support as well as purely technical assistance.

To build these resources, we will draw upon the tools and technologies from the Social Web (Reed et al 2004, Hoschka 1998), known also as Web 2.0 (O'Reilly 2004, 2005), which stresses collaboration and sharing among users of web-based services. The key observation is that the collective knowledge of users in a community is substantial, and if leveraged properly can benefit all users. We propose to develop a cluster of interrelated tools, including:

- **Wiki-enabled documentation** that would allow users to suggest corrections and improvements to the extensive documentation. The user community contains many experts with deep knowledge of specific subject areas and countries, and many are quite willing to share their expertise to help others.
- **Expert Q&A system** where users can pose specific queries. Volunteer experts can answer these questions by starting discussion threads; other users can comment on or clarify an answer, which generates better quality answers. These threads will then be archived and indexed by keywords, allowing users to search old queries before submitting a new one.
- **Specialized research forums** that bring together smaller groups of users with detailed knowledge on a topic. These forums would encourage research collaborations among scholars from diverse disciplines who otherwise might not interact.
- **Tools for sharing SAS, Stata, and SPSS code for data manipulation** developed by individual users that could also benefit others. Currently sharing is *ad hoc*, with no systematic match-making. The proposed approach will create a shared repository with a searchable directory.
- **Tools for sharing curricular materials** based on the same principles as code sharing. The software developed for code sharing can be substantially reused for the benefit of educators in identifying and sharing materials for teaching about IPUMS.
- **Expert recommendation system** for problems frequently encountered by users. The idea of this tool is to infer interests and requirements of users from their data requests and other activities, and then to recommend datasets, research forums, discussion threads from the expert Q&A, and code based on a "match-making" algorithm. This approach has been very successful in many domains and has been shown to improve user experience and effectiveness.

This is not a definitive list of community tools; rather, it is a starting point for planning and evaluation that must occur before we undertake software development. Each of these tools has shown substantial benefits in other environments, but they are new to social science research.

Design of the tools will be a collaboration with our users, informed by surveys, user feedback, and a Wiki-based discussion forum for each tool.

2. Aggregate data access. In recent years, multilevel analysis based on merging census microdata with census summary statistics for geographic areas has become a widely-used and powerful analytic approach. Constructing files suitable for multilevel analysis remains cumbersome. We propose to simplify multi-level analysis by allowing researchers to attach aggregate statistics to individual-level records as part of a data extract. Construction of the aggregate statistics files is discussed below in section 7, "New Data Products." Users will select their summary variables in much the same way that they now select individual-level variables, and they will specify the geographic level of measurement required for each variable. The system will then merge the summary data into microdata extracts. By automating the tedious process of constructing aggregate data and combining them with microdata, we will enrich both sources and stimulate a broad range of new research initiatives.

3. Variable search. One consequence of adding many new datasets is that the number of variables increases dramatically. This requires us to improve our technology for navigating metadata. Indeed, the number of variables is already outstripping the existing simple pick-list model for presenting available variables. Currently, the main IPUMS-International website has over 5,500 variables. As we add datasets, the number of variables will continue to grow. As the number of variables grows, the current approach will become increasingly unworkable. New tools will allow researchers to quickly identify the most suitable variables for a particular research problem. For example, users will be able to filter the variable list according to keywords or subject area; they will be able to reduce the list to only those variables appearing in every sample of interest or to expand it to include all variables in any selected sample; and they will be able to view simplified pick lists focusing on the most commonly requested variables, as determined through analysis of extract logs.

4. Online analysis. As noted above, in the coming year we will implement a simple online analysis system based on the SDA software. This system will make it possible to generate statistics without downloading microdata, decompressing the file, and analyzing the data using a software package. We regard this as an interim solution, since SDA has a poor user interface and is slower than optimal. In the next grant period, we envision a set of improvements that will make online analysis more usable.

The first priority is to improve the user interface to make it more intuitive. This would make an immediate contribution in several areas. Skilled researchers could use the system for quick reference and exploratory data analysis. Educators from primary to graduate school could use the system to teach data analysis without the costs, in time and money, of conventional statistical packages. To realize these goals, the new interface must incorporate the following features that are not available through existing online microdata analysis systems:

- *Full integration with the IPUMS-International metadata system.* Creating a system driven by existing metadata will reduce implementation and maintenance costs, prevent duplicated metadata, and minimize the potential for errors.
- *Information-rich environment for formulating queries and interpreting results.* Users should have immediate hypertext access to any variable-level information at every stage of analysis.

Thus, when users design their query or view their results, all variable labels should link to the appropriate IPUMS variable description, which in turn provides single-click access to the relevant questionnaires and instructions, universe, and other information.

- *Flexible and intuitive data manipulation.* This will include tools for recoding variables and constructing new variables based on the characteristics of family members (e.g., spouse's race).
- *Capability to save, retrieve, and modify analyses.* This feature is essential if the system is to become more than a novelty. It will allow users to replicate results from a previous session, and it encourages a cumulative approach to discovery that builds on prior analysis.
- *Elimination of software client.* Most previous online microdata analysis software requires a downloaded client to access their most advanced features (e.g., Nesstar, Data Ferrett, PDQ, Querylogic). By using new web technologies, described below, we can eliminate the need for a client, making the software easier to use and expanding the potential audience.
- *Convenient and documented output options.* We will provide an option to download tabular output in a readily transferable form (e.g., .csv file). Each analysis will be accompanied by customized documentation that describes the source data, universe, and any data manipulation used to create the table, such as subsetting, recoding, and weighting.

In addition to improving the user interface and providing more powerful capabilities than our existing system, we also plan to increase the speed of our online analysis. The SDA analysis engine is adequate for most individual samples. It is, however, too slow when analyzing merged files that may include tens or hundreds of millions of cases. Better alternatives are available, and we do not intend to reinvent the wheel. If possible, we will license high-speed tabulation software from another vendor, rather than building it ourselves.[3]

Implementation of web-based innovations. These software innovations—including variable search, online analysis, and Web 2.0 capabilities—will require us to exploit new web technologies. The classic synchronous model of web applications is poorly suited to the growing complexity of our data access system. Under this model, user actions in the interface trigger an HTTP request to the web server, which then returns an HTML page to the client. Ajax tools (Asynchronous JavaScript + XML) allow a much more flexible approach. The power of Ajax can be seen in several new web tools introduced during the past two years, such as Google Maps, Flickr, Orkut, and Gmail. Instead of rewriting the screen each time the user makes a request, these applications handle user requests asynchronously. The browser loads a javascript engine that renders the interface dynamically and handles communication with the server. When the engine needs to retrieve information from the server, it does so in the background, without interrupting the user's interaction with the application. Using this approach, the Gmail program allows users to browse through mailboxes that contain thousands of messages, filtering and selecting emails at least as rapidly as a desktop mail application. IPUMS-International requires the same kind of speed and flexibility.

[3] Two companies, PDQ.com and Querylogic, developed rapid tabulators for IPUMS data using Small Business Innovation Research funds from NIH or NSF. Both systems are extremely fast and are therefore better suited to the scale of IPUMS than are other online data analysis systems currently in use (e.g. Nesstar, Data Ferret, Virtual Data Center, or SDA). We are optimistic that one of these companies will have an interest in collaboration.

4. Geographic improvements

Over the past 18 months, we have made a concerted effort to add more geographic detail to the data series. For all samples where it is possible, we identify at least the second administrative level, typically referred to as municipality, county, or district. In most samples, this means the identification of any unit with a population in the most recent census of 20,000 or more. Smaller units are aggregated to achieve the necessary threshold.

In the short run—the 2008-2009 project period—we will continue to build on our geographic work and provide more geographic variables for more countries. In addition, where feasible we will provide scanned images of census maps that will illustrate the locations of the places identified. With additional time and funding in the 2009-2014 period, we will be able to make substantially greater progress on geography, identifying many more metropolitan areas, cities, and towns. In addition, we will add centroids for geographic places to the data, and construct migration variables describing the distance migrated and the direction of migration.

We also plan for the 2009-2014 period to develop electronic boundary files for the geographic units identified in the data. Such files are essential for applying GIS approaches to spatio-temporal analysis. Where possible, this project will build on boundary files produced by national statistical agencies, edited to maximize cross-national consistency and coded for compatibility with the microdata. Where pre-existing electronic boundary files are not available—or in cases where we cannot obtain a license to disseminate them—we will construct new files, capitalizing on software and methods developed at the Minnesota Population Center for the National Historical Geographic Information System.

We are also contemplating a more ambitious approach to IPUMS-International geography. Changing boundaries pose one of the most frustrating challenges to the spatial analysis of change between census years. Many analyses rely on interpolation to adjust the data to constant boundaries, but this tends to blur geographic differentials. Moreover, we usually lack the kinds of source data commonly used to generate interpolated statistics.

We have the potential to implement a far better approach. Most of the IPUMS-International data for developing countries include small-area identifiers, often down to the block level. We must suppress this information in public releases of the data because of the risk of respondent disclosure, but we can use it to construct new geographic units. In particular, it may be possible to create harmonized statistical areas with approximately 20,000 persons that are compatible over time, thus eliminating the intractable complexities associated with boundary changes. This would require information for each census on the physical location of the smallest geographic units in the data, and such metadata may be hard to locate for older censuses. In addition, to make the project practical, we would need to develop cost-effective methodology for constructing perhaps 100,000 statistical units in over 100 censuses. Accordingly, we are undertaking a pilot project to assess the feasibility of this approach.

5. Data quality assessments

One of the challenges of international comparative research using censuses or surveys is that data quality may vary substantially across countries and over time. Reports on data quality based on post-enumeration surveys or demographic analysis are irregular, and they do not exist at all for

many countries. There are few systematic evaluations of data quality that compare censuses across decades and between countries. A quarter-century ago, a National Research Council study evaluated 77 censuses conducted in developing countries between 1960 and 1980 (NRC 1981), but as Cleland (1996) points out, the estimates of underenumeration were made using widely varying methods and are not comparable across time or between countries. There have been a variety of regional and national studies of census quality (e.g., Anderson 2004, Chackiel 2001, Dorrington 2002, Luther 1983), but these studies seldom permit systematic comparisons of census reliability or coverage.

We plan to aid users in the evaluation of census quality by constructing several broadly comparable indicators of census quality. These will include:

- Age heaping. We will compute the standard indices (Whipple's Index, Myer's Index, PPI) to provide a comparable indicator of digit preference.

- Age-specific sex ratios. We will construct measures of sex ratio irregularities, such as the United Nations Age-Sex Accuracy Index (Siegel and Swanson 2004), and assess undercount of young, highly mobile working-age men.

- Cohort survival analysis. In the absence of reliable vital statistics for many countries, full demographic analysis is not feasible. Nonetheless, simple analysis of cohort survival between censuses can identify serious undercounts of infants, children, and young men, as well as overcounts of the elderly.

- Nonresponse. We will develop a comparable index of item nonresponse that summarizes the extent of missing data for a group of broadly available variables.

- Structural Error. In most censuses, data collection or processing problems lead to a certain proportion of households that do not conform to enumeration rules. For example, there are often households with multiple heads or individuals and family groups with no corresponding household record. IPUMS-International generally corrects these errors through imputation. We will provide statistics on the frequency and type of these structural errors in each census.

- Inconsistencies. We will develop a general index of internal inconsistencies in the data. These include, for example, inconsistency in the age of mothers and children (e.g., mothers younger than their children), inconsistency in marital status and relationship (e.g., spouses listed as never married), and so on. In all, the index will evaluate approximately 20 common inconsistencies.

These measures will not provide an indication of net undercount, but they will give a general sense of census quality and alert users to the samples that pose the greatest potential problems.

In addition to constructing these indicators of census quality, we will attempt to summarize results of post-enumeration surveys and demographic analysis carried out by national statistical offices, as well as published evaluations of censuses by individual scholars. In addition, we will solicit user assessments of census quality through the web community tools described above in section 3, "Cyberinfrastructure."

6. Data processing

The five topics discussed above—outreach and dissemination, data acquisition, cyberinfrastructure, geographic coding, and data quality assessment—represent a minority of the work involved in creating the database. The largest work component is data processing. As noted, by streamlining our processing procedures, metadata, and software, we have achieved unprecedented efficiencies, but data processing remains our biggest task. The following paragraphs briefly summarize the major processing steps required to add a dataset to IPUMS-International.

Language translation. We require that all key documents—most notably, the data dictionary, questionnaire, and enumeration instructions—be available in English before we commence data processing work.

Metadata preparation. Once the necessary documents are available in English, the first processing step is to document the input data. We receive data dictionaries in many formats and must transform this disparate documentation into a single systematic XML-encoded format easily readable by software. We also tag the text of census forms and enumeration instructions, write variable descriptions, and document sample design and census characteristics, such as *de jure* or *de facto* enumeration rule.

Data reformatting. Once the metadata describing the input files is complete, we can begin to transform the original data files into a standard format. Data come to us in a wide variety of formats; converting them to a standard format simplifies later stages of processing. Just as important, the reformatting stage involves running various diagnostics to discover problems. Data errors that affect the structural soundness of households and dwellings—for example, corrupted households consisting of mismatched individuals—must be corrected. During reformatting, we add some basic technical variables. These include both serial numbers (dwelling, household, and person number) and counts of households and persons within each dwelling. At the same time, we insert flags identifying households with multiple heads, no head, multiple spouses, duplicated records, and/or other conditions that may indicate faulty data.

Household substitution and sampling. In the majority of the datasets we have analyzed, a small fraction of dwellings have structural problems with no clear solution (Esteve and Sobek 2003). If there is no solution to a structural problem, then we mark the affected records as bad and substitute donors from other records in the dataset. We use whole-dwelling substitution, identifying appropriate predictor variables for each of the major types of dwellings in the data (usually multi-household, vacant, collective, and single-household private).

Confidentiality edits. In some cases, we receive fully anonymized samples from statistical offices; in other cases, the agencies implement some but not all of the necessary privacy measures before sending us the data; and in still other cases, we have virtually full information from the census (apart from actual names). Whenever necessary, we must implement statistical confidentiality edits approved by each national statistical office. We identify the lowest level of geography to be released and suppress all finer geographic variables. We also identify and suppress any other sensitive variables, and eliminate any technical variables that could be used to identify the record within the original data. In some instances, we must also eliminate other potentially identifying information, such as date of birth or full character string for occupation.

We then recode very small population categories for specific variables into larger groups (for example, grouping rare occupations with more common pursuits), and top- or bottom-coding some variables (for example, income). Finally, we randomize the sequence of dwellings within the smallest geographic unit identified in the data, so geography cannot be inferred from file position, and we randomly swap an undisclosed fraction of cases across geographic districts to add uncertainty about the origin of a particular record. Then we generate a new serial number to reflect the final ordering of the file.

Universe checks and data cleaning. Census forms often state the universe for a question, but the stated universe sometimes has no obvious correlates (in terms of a checkbox, clear skip pattern, or blank line for those "not in universe") on the form. In other cases, there are missing or errant values in the data. Finally, out-of-universe cases are often combined with logical zeroes or non-responses. We therefore empirically verify the universe of every input variable.

Integration. The culmination of IPUMS data processing is integration: designing variables for which the same codes mean the same things over time and across countries, and writing documentation that explains differences that persist in the final integrated variable. The goal of integration is to simplify analysis across time and space without losing any information. The standardization and documentation of the source materials described above greatly simplifies integration, but harmonizing variable coding remains an often-challenging logical puzzle. Although data integration involves intellectual work that no program can provide, we have developed software to aid in the logistics.

Documentation. When the integrated coding is complete, we expand all documentation for the integrated variable (such as the variable descriptions, codes and frequencies, and enumerator instructions) to account for the new sample and any changes in the codes. The comparability descriptions require particular care; research staff must decide what differences in census wording, concepts, or variable coding are worthy of mention in the integrated variable documentation. Both international and intra-national comparability need to be considered. Users will not be utterly dependent on our judgment, however: at a click they can examine the associated enumeration text for any integrated variable, or to examine the constituent unharmonized variables that served as input to the integrated version.

Data processing, 2006-2007

In the past 18 months, we more than doubled the number of samples in the data series. In May 2006, we added 19 samples from four Latin American countries and South Africa; in November 2006, we added 16 samples from seven European and developing countries; and in May 2007, we added 17 more samples from Argentina, Hungary, Israel, Palestine, Portugal, and Rwanda. These data releases occurred exactly as scheduled in our 2006 plan.

In the course of these data releases, we added approximately 100 integrated variables, yielding a current total of about 400. Some additions were variants of existing IPUMS variables needed to accommodate characteristics that differ across countries (e.g., geography, migration, and ethnicity); others were completely new integrated variables. As described above in section 3, "Cyberinfrastructure," we also added 5000 unharmonized variables. In addition, we upgraded the 28 original samples developed under the first IPUMS-International grant and standardized our

treatment of them to be consistent with current practices. This included, among other changes, major revisions to metadata and development of unharmonized variables.

Planned data releases, 2008-2009

We plan annual data releases for the last two funded project years, in May 2008 and May 2009. Each data release will include approximately 32 samples, bringing the total number of samples to about 144. This number substantially exceeds the 128 samples we estimated we would be able to produce under the approved budget for the project.

The May 2008 release will include samples for Austria, Canada, China, Iraq, Malaysia, Netherlands, Panama, the United Kingdom, and Venezuela. We have not yet finalized the list of countries to be included in the 2009 release, but we expect to include Indonesia, Pakistan, Bangladesh, Sudan, Egypt, and Thailand.

The 2008 and 2009 data releases will incorporate several new features and documentation, in addition to the items described above in the sections on cyberinfrastructure, geography, and data quality assessment.

Mother and father pointers. We will add constructed family interrelationship pointers for mothers, fathers, and spouses for all samples. During the first IPUMS-International grant period, we were conservative and limited parent-child links to persons under age 19. During this development period, we will explore the feasibility of removing that restriction.

Socioeconomic variables. Depending on the results of research over the previous year, we will add indicators of socioeconomic status to the standard constructed variables of the IPUMS. The most likely sources for such variables will be occupation, income, education, and dwelling characteristics.

Document data transformations. Most data transformations are documented in the integration tables, including how input data values and labels correspond to IPUMS-International values and labels. We will determine the most practical format for delivering this information to users on the web. The programming scripts that supplement the integration tables will be harder to render intelligible to users. We intend to invest considerable effort into documenting these transformations for internal purposes. In the process, we will keep in mind that these scripts will ultimately become public documentation. All modules of the data conversion program itself will also be made available, along with the metadata inputs for the program (such as the missing data allocation scripts).

Variance estimation. We will provide a discussion on our website of the impact of IPUMS-International sample designs for variance estimation. The discussion will offer advice to users on appropriate techniques and strategies for producing the valid estimates.

Planned data releases, 2009-2014

As anticipated in our original proposal, by the end of the current grant, we expect to have acquired and preserved far more data than we can process with our current resources. We have already received about 40 more samples than we will be able to process with the current grant. We are currently negotiating with about 30 countries for additional data, and we expect most of

those negotiations will be successful. As noted, we are also working with many countries to add older samples to the data series; in many cases, these data require extensive processing or tape recovery work. In a number of countries, we are also seeking larger or more detailed samples of the censuses we are already distributing. Taken together, these efforts could give us in 2009 a collection of unprocessed data that is almost as large as the body of released data.

Just as important, during the 2009-2014 period, we expect to acquire census data from the 2010 round of censuses for at least 80 countries. Updating the database with current data is an extremely high priority; however important the historical aspect of the database, if IPUMS-International is not kept up-to-date, it will lose salience as a tool for policy analysis. We will therefore have enough raw material to keep the data processing component of the project running at full capacity at least through 2014.

7. New Data Products, 2009-2014

For many developing countries, especially in Latin America and Africa, we have obtained entire enumerations or very high density sample data with full geographic identifiers. These data include over a billion records, and they open up the potential for entirely new approaches to spatial population research. Traditionally, small-area analysis has relied on tabulations of censuses produced by national statistical agencies. Such tables are extremely limiting, because investigators cannot assess the relationships among variables at the individual level. Moreover, the available tables invariably change from census to census both with respect to content and geographic boundaries, and they are usually incompatible from country to country.

We propose two initiatives that, we think, will have profound consequences for spatio-temporal analysis of human social dynamics. First, we propose creating restricted-use versions of the complete-count microdata files. Second, we propose creating an integrated set of aggregate summary tables that can be made public and will allow scholars to carry out multi-level analysis that spans national borders.

The restricted access complete count data would include the full geographic identifiers included in the original data. For the first time, analysts will be able to analyze individual-level interrelationships among variables at the finest levels of geography—in most cases, down to the block level. Scholars have never had access to data like this, so we do not know exactly how they will use it. It is certain, however, that this kind of small-area spatio-temporal microdata would stimulate the development of new methods of spatial analysis.

The biggest challenge posed by the complete-count data files is disclosure avoidance. We anticipate making the data available only through a network of restricted-access data enclaves that subscribe to rigorous confidentiality restrictions. To implement this plan, we will have to negotiate new agreements with each county that allow for controlled researcher access to the complete count data. In addition, we will have to undertake substantial data processing work, and the reformatting and imputation procedures we currently use will have to be modified to accommodate complete count files.

The aggregate data files will identify a consistent set of summary variables in the censuses that would be useful for multilevel analysis, and construct them from the microdata for the lowest

geographic level that can be publicly identified. At present, analysis of small-area data across census years and national boundaries is almost impossible because of the inconsistencies. Potential topics of analysis include residential segregation; immigrant and ethnic settlement patterns; suburbanization and urban sprawl; rural depopulation and agricultural consolidation; the identification of concentrated poverty; causes and levels of change in ecosystems; transportation; and environmental justice.

We are presently engaged in a project to create new summary files for the 1960 and 1880 United States censuses, and have formed a technical advisory group to council us on the highest-priority tables. This experience will serve us well when constructing aggregate data tables for developing countries. We will supplement these aggregated tables with summary variables at the level of region, country, and district compiled by the United Nations, the World Bank, and national statistical agencies describing economic characteristics (e.g., Gross Domestic Product per capita) and demography (e.g., life expectancy).

As described in the section 4, "Geographic improvements," we are contemplating development of harmonized statistical areas with approximately 20,000 persons that are compatible over time. Such harmonized geographies would be an important complement to both the complete-count microdata and the aggregate summary files. We could provide 10 percent microdata samples, complete-count restricted access microdata, aggregate data tables, and electronic boundary files, all with one standard geography that is consistent across time and consistently defined in each country. Taken together, these data sources would provide unprecedented opportunities to understand global-scale changes with fine resolution.

Conclusion

For the countries that have long had access to public use census microdata—the United States, Canada, and the United Kingdom—the data are an indispensable component of the infrastructure for social and economic research. For example, during the past two decades census microdata have been the most frequently used quantitative source in the pages of *Demography*, the leading journal for population research. For most countries, however, census microdata have never been widely available to researchers.

IPUMS-International is opening up new avenues for understanding social dynamics in country after country. The data releases of the past 18 months have made it feasible for the first time to do large-scale microdata analyses that span the globe. This report does not attempt to explain the significance of the research that is already being produced. In the appendix, however, we provide a bibliography of the early publications and papers using IPUMS-International which gives some idea of the extraordinary versatility and power of the data.

IPUMS-International is already the world's largest collection of publicly accessible microdata, and the project is still young. Over the next seven years, the database will expand dramatically, and we will develop new tools and metadata that will allow researchers to fully exploit this complex large-scale resource. Although we have not fully described all the improvements we hope to undertake over the next seven years, we hope this report at least gives a sense of the extraordinary opportunities and challenges that lie ahead.

References

Anderson, Barbara. 2004. "Undercount in China's 2000 Census in comparative perspective." Ann Arbor, Michigan, University of Michigan, Institute for Social Research, Population Studies Center, PSC Research Report No. 04-565.

Chackiel, Juan. 2001. "Censuses in Latin America: new approaches. Symposium on Global Review of 2000 Round of Population and Housing Censuses: Mid-Decade Assessment and Future Prospects. Statistics Division, Department of Economic and Social Affairs, United Nations Secretariat, New York, 7-10 August 2001

Cleland, John. 1996. "Demographic Data Collection in Less Developed Countries 1946-1996 *Population Studies* 50: 433-450.

Dorrington, R. 2002. "Did they jump or were they pushed? An investigation into the apparent undercount of whites in the 1996 South African census." *South African Journal of Demography*. 8(1):37-46.

Hoschka, Peter. 1998. "The Social Web Research Program: Linking people through virtual environments." Web document accessed 2/24/07 at http://www.fit.fhg.de/~hoschka /Social%20Web.htm.

Luther, Norman Y. 1983. "Measuring changes in Census coverage in Asia" *Asian and Pacific Census Forum*, 9(3):1-11.

National Research Council. 1981. *Collecting Data for the Estimation of Fertility and Mortality*. Washington, D.C.: National Academies Press.

O'Reilly, Tim. 2005 "What is Web 2.0?: Design Patterns and Business Models for the Next Generation of Software". 30 September. Web document accessed 2/24/2007 at http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.

O'Reilly, Tim (Organizer). 2004. First Annual Web 2.0 Conference. 5-7 October, San Francisco, CA. Web document accessed 2/24/2007 at http://www.web2con.com /web2con/.

Reed, Drummond, Marc Le Maitre, Bill Barnhill, Owen Davis, and Fen Labalme. 2004. "The Social Web: Creating an Open Social Network with XDI." *PlaNetwork Journal*. Web document accessed 2/24/07 at http://journal.planetwork.net/article.php?lab=reed0704.

Siegel, Jacob S. and David A. Swanson. 2004. *The Methods and Materials of Demography, Second Edition*. London: Elsevier.

United Nations Economic Commission for Europe. Conference of European Statisticians. 2007. Final Guidelines on Managing Statistical Confidentiality and Microdata Access: Principles and Guidelines of Good Practice. Geneva: Publication No. E.07.II.E.7 . http://www.unece.org/stats/documents/tfcm.htm [see case study 23, pp. 98-103].

A REVIEW OF IPUMS-INTERNATIONAL

Dennis Trewin Statistical
Consultant

1. Terms of Reference

IPUMS-International is a project to inventory, preserve, harmonise and disseminate census microdata from around the world. Use is restricted to scholarly purposes. It is a collaboration of the Minnesota Population Center (MPC), national statistical agencies, international data archives and experts from participating countries. Major funding for IPUMS-International is provided by the National Science Foundation and the National Institutes of Health.

The Terms of Reference for this study are as follows.

"The goal of IPUMS-International is perfection in the following sense:

a) total satisfaction of our statistical agency partners that we have implemented every detail of the memorandum of understanding regarding both archiving and dissemination
b) complete satisfaction of the researchers regarding the integrated data and documentation

The study will identify weaknesses and lapses so that IPUMS-International can improve its procedures. This will provide an additional layer of protection for official statisticians as well as trust for the public."

To achieve the last objective it will be necessary for this review report to be a public document. It has been written with this in mind.

A number of information sources were used to undertake this review.

(i) The role of 'archiving' services like IPUMS-international was discussed at the 2007 Session of the United Nations Statistical Commission. I was the author of the main discussion paper and, at the time, an Interregional Adviser on National Statistical Systems for the United Nations. As a consequence I also participated in a number of 'corridor' discussions on the topic.

(ii) A number of meetings with participating countries were held in conjunction with the 2007 Session of the International Statistical Institute. I was able to participate in some of those discussions as well have bilateral discussions with country experts.

(iii) I was able to read a number of papers and reports relevant to IPUMS-International.

(iv) The IPUMS-international procedures are well documented facilitating a review of their adequacy.

(v) I have had extensive discussions with the staff of the Minnesota Population Center (MPC).

2. Why am I qualified to undertake this review?

Until January 2007 I was Australian Statistician responsible for the Australian Bureau of Statistics. One of my achievements was the extension of microdata services to researchers whilst maintaining public trust and abiding by the conditions outlined in the legislation governing microdata access. During this time I was asked by the Conference of European Statisticians to chair a Task Force to produce guidelines on good practice on the release of microdata and the protection of confidentiality. These were published in 2007 as "Managing Statistical Confidentiality and Microdata Access: Principles and Guidelines on Good Practice" (CES, 2007). I produced a variant of these guidelines as a Discussion Paper for the 2007 session of the United Nations Statistical Commission. These were discussed and adopted as international guidelines.

3. Objectives of IPUMS-International

The objective of IPUMS-International is to make cross-national census microdata readily accessible and usable. The project facilitates comparative international research based on pooled microdata. A harmonised composite coding system for variables allows easy comparisons across time and countries. Extensive documentation aids in the interpretation of data.

4. Services Provided

IPUMS-International provides a range of services both to data providers and those wishing to access the data users. It is important not to under-emphasise the services to data providers. For many it is a major motivation for belonging to IPUMS.

Some of the key services provided to data providers include:-

(i) Microdata creation services – assistance in creating a confidentialised microdata file in a form that can be used by researchers.

(ii) Data archiving services – a copy of a country's data becomes available off site at MPC; it is recommended practice that key data sets be archived off site but many countries do not have the facilities or resources for doing their own archiving. There are several examples of countries recovering lost data (eg through natural disaster or mismanagement) by using the IPUMS service.

(iii) Data recovery services – IPUMS expertise or contractors working with IPUMS has been able to be used to recover data where the storage medium (eg magnetic tapes) have been damaged.

(iv) Seminar services – international seminars on various topics are a very effective way of sharing knowledge.

(v) Documentation services – the documentation produced by IPUMS can often be very useful for the country providing the data.

(vi) Improved data access – countries can use IPUMS tools and facilities to generate tables and other statistical material from their own data.

IPUMS-International is considering an increasingly range of powerful on-line facilities to allow improved researcher access to Population Census microdata. These may be very useful to some users although the majority of users will still prefer simply to have data sets they can download into their own software facilities. But there will be an important sub-set of users who will not have access to such facilities and would prefer on-line facilities.

For data users, the prime service is access to a rapidly growing set of microdata that has been 'integrated'. The process of integration greatly facilitates comparisons across countries – it also facilitates comparisons across time.

Among other things, it supports cohort analysis. It enables cohorts to be tracked across successive censuses. Although not as powerful as longitudinal analysis, research studies have shown that cohort analysis provides at least half the power of longitudinal analysis and can be far easier to implement.

We should not forget the benefits from simply having the data available for research use. For many researchers, it may be the only way to access some data sets especially those from developing countries.

It should also be seen in the context of international trends in making access to more detailed data increasingly available but in a way that its confidentiality is protected. The demand is there – technology has made detailed data easier to access and it has also made evidence based decision making increasingly realistic.

5. Trends in the Provision of Microdata

The 2003 Conference of European Statisticians was the first occasion the heads of National Statistical Offices collectively considered the question of microdata. Importantly they agreed that supporting research is an important activity for National Statistical Offices (NSOs) and that most NSOs could do more to satisfy these needs. Doing more included providing improved access to microdata. They commissioned work which led to the publication of Guidelines on Confidentiality and Microdata Access (CES 2007).

6. Broad Conclusion

IPUMS-International provides a range of deeply appreciated services with rapidly increasing demand particularly in the United States. It also has the potential for even greater use internationally especially by the international agencies. It could become one of the most important global statistical assets.

Without question IPUMS International meets the four Core Principles outlined in CES (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 of international statistical data sets.

7. How does IPUMS-International work?

A more detailed description is given as Annex 1.23 in CES (2007)

The core processes are:

(i) obtaining samples of microdata and supporting metadata from participating countries. This does not happen before a Memorandum of Understanding outlining the conditions of release is signed at a senior level by both the University of Minnesota and the participating countries (usually a representative of the National Statistical Office).

(ii) The Memorandum of Understandings are cleared with the legal and accounting departments at the University of Minnesota. The standard contract was developed in close liaison with the General Counsel at the University and he also approves deviations from the standard. The accounting department signs each individual memorandum so it is an arms length process from MPC.

(iii) Before this data is released for researcher access, it is confidentialised, harmonised (to facilitate comparisons across countries and time), and structured to enable efficient access. Documentation is also structured to enable easier access.

(iv) Researcher access is limited to scholarly purposes. Before being given approval to access the data, researchers must apply for access. The registration process requires them to agree to a range of conditions. It also points out that breaches of these conditions make them liable to a range of penalties including withdrawal of service or possible legal sanctions.

(v) There is an expectation that they will notify IPUMS–International of publications, reports, etc that result from their access to IPUMS data.

Data is often confidentialised before it is provided to IPUMS-International. If it is not, MPC will take further steps such as constraining the level of geographic and occupation detail. It also uses data swapping techniques as a further layer of confidentiality protection.

Of course further protection is provided by the steps IPUMS-International will take if there are breaches of the conditions of access.

8. Feedback from Data Providers

The United Nations Statistical Commission discussed data repositories in March 2007. Some of the main conclusions of that discussion were:

(i) countries should retain ownership of their submitted data and control over its access;

(ii) there should be greater transparency in policies governing the release of data; and

(iii) there should be consultation between countries and the depositories on the arrangements for granting access to data.

My other main source of feedback from countries providing data to IPUMS International was at the Seminar held in Lisbon 2007. As well as the discussions held during the Seminar, it was an opportunity to have bilateral discussions with participants. Some of the key points made in these discussions were:

(i) They would like more feedback on researchers accessing their data or research that resulted from this access,

(ii) The IPUMS-International services are highly appreciated by National Statistical Offices (NSOs) especially the archiving, documentation and data recovery services;

(iii) There seemed to be strong support for an on-line facility available through the web; and

(iv) They would like to know more about practical statistical methods relevant to microdata such as confidentiality methods, data management and analytical skills.

The key message from them is one of strong support for IPUMS-International but with more feedback on research and researchers. This is discussed later.

Points (ii) and (iii) above are mainly aimed at repositories other than IPUMS-International. Data providers generally regarded the practices of IPUMS-International as being better than those of other repositories.

9. Feedback from Data Users

This is based purely on advice from the MPC. But these views are “evidence based”. They were compiled from an analysis of feedback MPC had received from data users.

The key message was one of overwhelming support for the program. The main request was simply for more data – more countries and more censuses. They emphasised the importance of harmonisation – it made analysis across countries and across time far more meaningful.

There was support for adding a sampling function to allow smaller data sets to be chosen. The data sets for some countries are very large and some users would prefer to work with smaller data sets.
There also seemed to be a demand from some users to more easily drop the data from IPUMS – International into GIS systems.

There were also requests for more workshops on IPUMS–International so that users could better understand the data sets and how to use them.

10. Comparison with International Guidelines on Microdata

The Principles are outlined in CES (2007). They are as follows.

1. It is appropriate for microdata collected for official statistical purposes to be used for statistical analysis to support research as long as confidentiality is protected.

2. Microdata should only be available for statistical purposes.

3. Provision of microdata should be consistent with legal and other necessary arrangements that ensure the confidentiality of released microdata is protected.

4. The procedures for research access to microdata, as well as the uses and users of microdata, should be transparent and publicly available.

With respect to the first principle, IPUMS-International protects confidentiality through the procedures described in Section 7. The limitation that IPUMS-International is only available for scholarly purposes, and the proposed uses are checked before approval to access is given, satisfies Principle 2. The Memorandum of Understandings signed both by the University of Minnesota and the country providing the data are an important way of satisfying Principle 3. The other important element is the undertakings signed by the researchers. The procedures for research access are available on the IPUMS-International web site thereby satisfying that element of Principle 4. The names of the researchers are not made available because the numbers are very large and it would be impractical. The web site does contain a Bibliography of papers published which are based on data obtained from IPUMS-International.

IPUMS-International is in full compliance with these Guidelines. For this reason, it was chosen as a Case Study on Good Practice on repositories in CES (2007).

11. Conclusions

1. There has been considerable growth in the demand for microdata for statistical analysis and research. This is reflected in IPUMS-International where usage has doubled over the last 12 months. There is no reason not to expect this type of growth to continue. Census microdata is a valuable data set and the harmonisation processes incorporated by IPUMS-International add considerably to value of these data sets.

2. Despite this rapidly growing demand, there is scope to improve awareness. Usage is dominated by students, particularly from USA, and other users from USA. There is scope for much greater use outside USA and by international organisations. Word of mouth (by satisfied users) is often the most effective way of increasing awareness.

3. IPUMS-International provides an excellent service:

    (i) to researchers and analysts by providing easy access and well documented data sets;

    (ii) to national statistical offices by (i) providing means of accessing their Census microdata, (ii) assisting them to produce good quality documentation of their censuses, and (iii) a Census data archiving services; and

    (iii) to some national statistical offices by helping them recover data where storage medium may have been damaged.

4. There is an ambitious program of development – not all proposed developments may be possible and some prioritisation may be necessary. The key user demand is for more Census microdata sets (especially from additional countries).

5. There is a high level of trust in the procedures adopted by IPUMS-International, both by data users and data producers.

6. 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.

7. The intention to encrypt data sets being moved to and from IPUMS was noted. This would be a very positive step towards securing the data.

8. Confidentiality is well managed but there are some areas of risk that may have been underestimated – these mainly exist (i) where data sets for regions (down to 20,000 people) can be recognised, and (ii) where there are relatively large samples (eg 10%) of relatively small countries. The proportion of unique households, even after removing geographic details and combining ages into 5 year age groups is about 25% for a medium sized country like Australia (20 million). This is from only examining the demographic characteristics of households. The data swapping technique used by IPUMS-International introduces a large degree of uncertainty into whether a record is unique or not but the increasing availability of demographic data sets from other sources (often private sector) means that confidentiality can never be absolute.

9. Consequently, there has to be some reliance on trust and/or undertakings which are enforceable. IPUMS-International pursues both approaches. Access is restricted to scholarly research and all users must sign an undertaking with full knowledge of the consequences of non-compliance. I support this approach.

10. But if the undertaking is to have impact, breaches must be seriously treated. Enforcement must be tested in cases of breaches. Depending on the seriousness of the breach, possible responses might be:

    (i) warnings to the researcher,

    (ii) withdrawal of service to the researcher, and possibly the institution for period of time,

    (iii) pursuance of charges of academic misconduct, and

    (iv) pursuance of legal options Assurances were made by IPUMS-International that these types of responses would be followed if necessary. But the undertakings will only be meaningful if any breaches are taken seriously.

11. The confidentiality risk could be reduced by limiting the amount of geographic and/or demographic detail available on the microdata sets at most risk. The amount of detail for more sensitive variables might also be reduced. Of course, such steps also reduce the usefulness of data sets. The confidentiality conditions might be reinforced to users in such cases.

12. Some countries will never be able to provide publicly accessible microdata to IPUMS-International because of legal or other constraints. Other approaches may need to be considered for such cases. Several possibilities exist.

(i) Countries may require individual approval of each access to their microdata set to be done by them (but IPUMS-International) would provide the access after approval has been received).

(ii) As for (i) but the countries providing the access.

(iii) Only providing access through the countries' own remote access facilities but they would be prepared to work with IPUMS-International to allow their data sets to be harmonised with the microdata sets of other countries.

Although not the preferred approach, these alternatives need to be considered if IPUMS-International is to become a truly global product. For these approaches to be effective, there has to be an agreed turnaround time to requests for approval. It is noted that (i) and (ii) make cross-national studies more difficult.

13. Feedback to countries on approved researchers is limited. Ways of notifying them of approved researchers should be considered. Their particular sensitivity is to researchers who might compute alternative estimates to official statistical estimates. This is most likely for developing countries – it is most likely to be done by international organisations. Perhaps notification could be limited to researchers from these organisations especially given the large number of researchers accessing the data sets.

14. Feedback on research is limited. A bibliography exists on the IPUMS-International web site. You appear to be able to search this data base by country. Perhaps this facility might be more widely promoted to participating countries.

15. There are some other initiatives that might be undertaken.

(i) A module on ethical behaviour should be included in training workshops and the like.

(ii) All applicants for access should be able to successfully complete a short questionnaire (8 to 10 questions say) which asks about their behaviours in certain circumstances.

(iii) The agreements to access are not very prominent on the IPUMS-International web site. There should be a reference on the home page where a single click would provide the reader with the conditions.

16. Countries participating in IPUMS-International want to learn more about good practice in areas such as metadata and data management, practical methods for data confidentiality.

17. IPUMS-International provides an outstanding service to both data producers and data users. These services are not as widely known as they should be among the international statistical community. Considerable effort has gone into improving awareness but more effort needs to be targeted at strategic organisations. These include the United Nations Statistics Division, World Bank, UNFPA, Eurostat and the UN Regional Organisations in Africa, Asia, the Middle East and the former Soviet Union.

12. Recommendations

IPUMS-International is a valuable and trustworthy microdata service. It meets the fundamental principles of good practice with respect to confidentiality and microdata. Consequently, my recommendations are limited.

In terms of the confidentiality arrangements, there are some minor adjustments that could be made in terms of improving the exposure of applicants and researchers to their obligation (See conclusions 11 and 15).

In the event of a breach appropriate action must be taken and the consequences of that action made aware to all users of the services (but in a way that does not identify the person making the breach). Checks should be made of published outputs from time to time will provide some assessment of whether there has been any inappropriate use of microdata (eg reference to individual cases).

Data providers could be given better feedback on researchers accessing their data sets (see conclusion 13). Efficient ways of doing this should be explored.

Awareness of IPUMS-International services could be improved by finding ways of better engaging with the relevant international agencies. This would require an assessment of the best way of approaching these organisations.

Alternative ways of incorporating the data for some countries needs to be considered (see Conclusion 12).

Finally, it is clear that IPUMS-International is on a strong growth path. The number of data sets that are part of IPUMS-International is expected to increase rapidly. Likewise the number of researchers using the service is expected to grow rapidly. There should be a specific position of how well positioned the MPC is to scale up to meet this level of expected growth.

13. Reference

CES(2007), “Managing Statistical Confidentiality and Microdata Access: Principles and Guidelines on Good Practice”, published on Conference of European of Statisticians web site (www.unece.org/stats)

Dennis Trewin
November 2007

IPUMS-International Publications

1. Journal articles and working papers

A'Hearn, B., J. Baten and D. Crayen. 2006. "Quantifying Quantitative Literacy: Age Heaping and the History of Human Capital," *Universidad Pompeu Fabra Economic Working Paper No. 996*, 2006.

Anriquez, Gustavo. 2007. "Long-Term Rural Demographic Trends," *ESA Working Paper No. 07-19*, FAO, Rome. 2007. (Background Paper to the World Bank's *2008 World Development Report.*)

Aydemir, Abdurrahman and George J. Borjas. 2006. "A Comparative Analysis of the Labor Market Impact of International Migration: Canada, Mexico, and the United States," *NBER (National Bureau of Economic Research) Working Paper No. W1327*, 2006.

Barbieri, A.F., R.L.M. Montemór and R.E. Bilsborrow. 2007. "Towns in the Jungle: Exploring Linkages between Rural-Urban Mobility, Urbanization and Development in the Amazon," in: *Workshop on Urban Population, Development and Environment Dynamics, 2007, Nairóbi, Kenya*. Paris : IUSSP/CICRED/PERN, 2007.

Bleakley, Hoyt. 2007. "Malaria in the Americas: A Retrospective Analysis of Childhood Exposure," BREAD (Bureau for Research and Economic Analysis of Development) *Working Paper No. 142,* Harvard University, 2007.

Cerda, Rodrigo. 2007. "Cambios Demográficos: Desafíos y Oportunidades de un Nuevo Escenario, " *Facultad de Ciencias Económicas y Administrativas, Pontificia Universidad Católica de Chile, Serie de la agenda pública, nº 11*, Santiago, Chile, October 2007.

Cruces, Guillermo and Sebastian Galiani. 2003. "Causality, Internal and External Validity, Childbearing and Female Labor Supply," *William Davidson Institute Working Papers number 626*, 2003.

Cruces, Guillermo and Sebastian Galiani. 2007. "Fertility and Female Labor Supply in Latin America: New Causal Evidence," *Labour Economics*, vol. 14 (2007), p. 565-573.

Dahl, Gordon B. and Enrico Moretti. 2004. "The Demand for Sons: Evidence from Divorce, Fertility, and Shotgun Marriage," *NBER Working Paper 10281*, January 2004.

Demont, Floriane and Patrick Heuveline. 2008. "Household Structure in post-Khmer Rouge Cambodia," *Journal of Population Research, Special Issue: New Approaches to Household Diversity and Change, Australian Population Association, Canberra, Australia* (forthcoming).

Demont, Floriane. 2008. "A travers l'évolution des structures familiales: Un retour sur l'Histoire récente de la péninsula indochinoise," *Les Cahiers Québecquois de Démographie*. Forthcoming.

De Vos, Susan and Luisa Schwartzman. 2008. "Using Union Status or Marital Status to Study the Living Arrangements of Elderly People" *Research on Aging*. Forthcoming.

Docquier, Frederic, B. Lindsay Lowell and A. Marfouk. 2007. "A Gendered Assessment of the Brain Drain," *Discussion Paper 2007-45*. Department of Economics, Catholic University of Louvain.

Docquier, Frederic and Abdeslam Marfouk. 2004. "Measuring the International Mobility of Skilled Workers (1990-2000): Release 1.0 (August 19, 2004)," *World Bank Policy Research Working Paper No. 3381*.

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

Esteve, Albert and Robert McCaa. 2007. "Educational Homogamy in Mexico and Brazil, 1970-2000: Guidelines and Tendencies," *Latin American Research Review*, vol. 42 (2007), p. 56-85.

Feliciano, Cynthia. 2005. "Educational Selectivity in U.S. Immigration: How Do Immigrants Compare to Those Left Behind?" *Demography*, vol. 42 (February 2005), p. 131-152.

Feliciano, Cynthia. 2008. "Gendered Selectivity: U.S. Mexican Immigrants and Mexican Non-Migrants, 1960-2000," *Latin American Research Review*, vol. 43 (2008), p.139-160.

Foldvari, Peter and Bas van Leeuwen. 2005. "An Estimation of the Human Capital Stock in Eastern and Central Europe," *Eastern European Economics*, vol. 32 (2005), p. 55-68.

Freeman, Richard B. 2006. "Does Globalization of the Scientific/Engineering Workforce Threaten U.S. Economic Leadership?" *Innovation Policy & the Economy*, vol. 6 (2006), p. 123-157.

Fussell, Elizabeth, Anne H. Gauthier and Ann Evans. 2007. "Heterogeneity in the Transition to Adulthood: The Cases of Australia, Canada, and the United States," *European Journal of Population*, vol. 23 (2007), p. 389-414.

Gonçalves Bueno Figoli, Moema. 2006. "Evolution of Education in Brazil: An Analysis of Educational Rates Between 1970 and 2000 According to Highest Grade Concluded," *Revista Brasileira de Estudos de População*, vol. 23 (January/June 2006), Sao Paulo.

Gupta, Neeru, Pascal Zurn, Khassoum Diallo and Mario R. Dal Poz. 2003. "Uses of Population Census Data for Monitoring Geographical Imbalance in the Health Workforce: Snapshots from Three Developing Countries," *International Journal for Equity in Health*, vol. 2 (2003), p. 1-10.

Huber, Susanne and Martin Fieder. 2007. "Strong Influence of Month of Birth on the Reproductive Performance of Vietnamese Women: Population Census Study," in process.

Hunt, Gary L. and Richard E. Mueller. 2004. "Canadian Immigration to the U.S., 1985-1990: Estimates from a Roy Selection Model of Differences in Returns to Skill," *Review of Economics and Statistics*, vol. 86 (2004), p. 988-1007.

Hunt, Gary L. and Richard E. Mueller. 2006. "The Migration of Highly Skilled Individuals Within and Between Canada and the United States," *Human Resources and Social Development Canada/Industry Canada/Social Science and Humanities Research Council, Skills Research Initiative, Working Paper 2006 D-14*.

Hunt, Gary L. and Richard E. Mueller. 2008. "Taxes and North American Labour Migration by Skill Level: Canada and the U.S. 1995-2001," in process.

Jones, J.H. and B.D. Ferguson. 2008. "The Marriage Squeeze in Colombia, 1973-2005: The Role of Excess Male Death," *Social Biology*. Forthcoming.

Klein, Herbert and Francisco Vidal Luna. 2004. "Sources for the Study of Brazilian Economic and Social History on the Internet," *Hispanic American Historical Review*, vol. 84 (2004), p. 701-716.

Lagakos, David. 2007. "Explaining Cross-Country Productivity Differences in Retailing," Working Paper, UCLA, Department of Economics.

Lagakos, David. 2007. "Superstores or Mom and Pops? Market Size, Technology Adoption, and TFP Differences," Working Paper, UCLA, Department of Economics.

Lam, David and Letícia Marteleto. 2008. " Family Size of Children and Women during the Demographic Transition," *Population and Development Review*, June 2008. Forthcoming.

Lam, David and Letícia Marteleto. 2006. "Stages of the Demographic Transition from a Child's Perspective: Family Size, Cohort Size, and Children's Resources," *Population Studies Center Research Report 06-591*, University of Michigan, January 2006.

Lambert, Paul, Vernon Gayle, Larry Tan, Ken Turner, Richard Sinnott, and Ken Prandy. 2007. "Data Curation Standards and Social Science Occupational Information Resources," *International Journal of Digital Curation*. vol. 2, (2007), p. 73-91.

Lutz, Wolfgang, Anne Goujon, Samir K.C. and Warren Sanderson. 2007. "Reconstruction of Populations by Age, Sex and Level of Educational Attainment for 120 Countries for 1970-2000," *IIASA Interim Report IR-07-002*. IIASA: Laxenburg.

McCaa, Robert, Rodolfo Gutiérrez and Gabriela Vásquez. 2000. "La Mujer Mexicana Económicamente Activa: Son Confiables los Microdatos Censales? Una Prueba a Través de Censos y Encuestas. México y los Estados Unidos, 1970-1990." *Papeles de Población*, vol. 6 (2000), p. 151.

McCaa, Robert and Steven Ruggles. 2002. "The Census in Global Perspective and the Coming Microdata Revolution," *Scandinavian Population Studies*, vol. 13 (2002), p. 7.

McCaa, Robert. 2002. "Unlocking the Census and Making it Usable: The IPUMS-International Consortium," *Paris-21 Newsletter*, vol. 1 (2002), p. 9.

McCaa, Robert. 2003. "El Calli Nahua del Mexico Antiguo: Hogar, Familia y Genero," *Revista de Indias*, vol. 63 (2003), p. 79-104.

McCaa, Robert and Albert Esteve. 2004. "La Integración de los Microdatos Censales de América Latina: el Proyecto IPUMS," *Estudios Demográficos y Urbanos*, vol. 58 (2004), p. 37-70.

McCaa, Robert and Albert Esteve. 2006. "IPUMS-Europe: Confidentiality Measures for Licensing and Disseminating Restricted Access Census Microdata Extracts to Academic Users," *Monographs of Official Statistics: Work Session on Statistical Data Confidentiality*, Luxembourg: Office for Official Publications of the European Communities, 2006, p. 37-46.

McCaa, Robert, Albert Esteve, Steven Ruggles and Matt Sobek. 2006. "Using Integrated Census Microdata for Evidence-based Policy Making: the IPUMS-International Global Initiative," *African Statistical Journal*, vol. 2 (2006), p. 83-100.

McKenzie, David, John Gibson, and Steven Stillman. 2006. "How Important Is Selection? Experimental vs. Non-Experimental Measures of the Income Gains from Migration," *Institute for the Study of Labor Discussion Paper IZA-2087*, April 2006.

McKenzie, David J. 2008. "A Profile of the World's Young Developing Country International Migrants," *Population and Development Review,* vol. 34 (2008), p. 115-135

Michaels, Guy. 2006. "The Division of Labor, Coordination, and the Demand for Information Processing," *Centre for Economic Performance, London School of Economics, Discussion Paper no. 0811*.

Mishra, Prachi. 2006. "Emigration and Wages in Source Countries: Evidence from Mexico," *International Monetary Fund Working Paper No. 06/86*, 2006.

Mueller, Richard E. 2006. "What Happened to the Canada-U.S. Brain Drain of the 1990s? New Evidence from the 2000 U.S. Census," *Journal of International Migration and Integration*, vol. 7 (2006), p. 167-94.

Oreopoulos, Philip. 2003. "Do Dropouts Drop Out Too Soon? International Evidence from Changes in School-Leaving Laws," (December 2003). *NBER Working Paper No. W10155*.

Rendall, Michael S. and Berna M. Torr. 2007. "Emigration and Schooling among Second-Generation Mexican-American Children," *RAND Working Paper WR-259*.

Ruggles, Steven, Miriam King, Deborah Levison, Robert McCaa and Matthew Sobek. 2003. "IPUMS-International," *Historical Methods*, vol. 36 (2003), p. 60-65.

Ruggles, Steven and Misty Heggeness. 2008. "Intergenerational Families in Developing Countries," *Population and Development Review.* June 2008. Forthcoming.

Ruggles, Steven. 2008. "Reconsidering the Northwest European Family System: Living Arrangements of the Aged in Comparative Historical Perspective," *Minnesota Population Center Working Paper 2008-2*. Under consideration by *Population Studies*.

Ruijs, Arjan. 2007. "Welfare Distribution Effects of Water Pricing Policies," *Fondazione Eni Erico Mattei Working Papers no. 151*. Available at SSRN: http://ssrn.com/abstract=1017476

Sandu, Dumitru. 2007. "Community Selectivity of Temporary Emigration from Romania," *Romanian Journal for Population Studies.* Under review.

Sassler, S. L. 2006. "School Participation among Immigrant Youths: The Case of Segmented Assimilation in the Early 20th Century, " *Sociology of Education*, vol. 79 (2006), p. 1-24.

Schouten, Barry and Marc Cigrang. 2004. "Remote Access Systems for Statistical Analysis of Microdata," *Statistics and Computing*, vol.13 (2004), p. 381-389.

Sinke, Suzanne M. 2006. "Gender and Migration: Historical Perspectives, " *International Migration Review*, vol. 40 (March 2006), p. 82-103.

Soloveichik, Rachel. 2007. "Family Transfers in Rural Mexico: An Application to Risk Sharing and Labor Supply Elasticity," in process.

Thomas, Wendy L. and Robert McCaa. 2003. "Archiving Census Documentation and Microdata: Preserving Memory, Increasing Stakeholders," *Notas de Poblacion*, vol. 29 (2003), p. 303-320.

Tsang, Daniel C. 2004. "Reflections on a Search for Social Science Data in Vietnam." *IASSIST Quarterly*, Spring 2004, p. 18-21.

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

Yoshioka, Hirotoshi. 2006. "A Q-Analysis of Census Data: Intra-Household Income Allocation and School Attendance in Chiapas, Mexico," *Quality and Quantity*, vol. 40 (2006), p. 1061-1077.

Yu, Eunice and Jianguo Liu. 2007. "Environmental Impacts of Divorce". *Proceedings of the National Academy of Sciences of the United States*. vol. 104, (2007), p. 20629-20634.

2. Books and Book Chapters

Bradley, Don E. and Charles F. Longino, Jr. 2008. "Geographic Mobility and Aging in Place," in Peter Uhlenberg, ed., *International Handbook of the Demography of Aging*. Forthcoming.

Chauvel Louis. 2006. “Social Generations, Life Chances and Welfare Regime Sustainability, “ in Pepper D. Culpepper, Peter A. Hall and Bruno Palier (eds.), *Changing France, The Politics that Markets Make*. Hampshire: Palgrave Macmillan, 2006, p. 150-175.

Chauvel Louis. 2006. “Générations Sociales, Perspectives de Vie et Soutenabilité du Régime de Protection Sociale,” in Pepper D. Culpepper, Peter A. Hall and Bruno Palier (eds.), *La France en Mutation 1980-2005*, Presses de Sciences Po, Paris, 2006. p.157-196.

Docquier, Frederic and Abdeslam Marfouk. 2006. “International Migration by Educational Attainment, 1990-2000,” in Caglar Ozden and Maurice Schiff (eds.), *International Migration, Remittances, and the Brain Drain.* Washington, DC: The World Bank and Palgrave McMillan, 2006, p. 151-200.

Fussell, Elizabeth. 2005. “Measuring the Early Adult Life Course in Mexico: An application of the Entropy Index,” in Ross MacMilllan (ed.), *The Structure of the Life Course: Standardized? Individualized? Differentiated?*Advances in Life Course Research, Vol. 9. Elsevier: JAI Press, 2005, p. 91-124

Hall, Patricia Kelly, Robert McCaa and Gunnar Thorvaldsen, eds. 2000. *Handbook of International Historical Microdata for Population Research*, Minneapolis: Minnesota Population Center and International Microdata Access Group, 2000.

Hunt, Gary L. and Richard E. Mueller. 2004. “International and Interregional Migration in North America: The Role of Returns to Skill,” in Barbara Messamore (ed.), *Canadian Migration Patterns from Britain and North America*. Ottawa: University of Ottawa Press, p. 229-44.

Le, C.N. 2007. *Asian American Assimilation: Ethnicity, Immigration, and Socioeconomic Attainment*. New York: LFB Scholarly.

Lloyd, Cynthia B., ed. 2005. *Growing Up Global: The Changing Transitions To Adulthood In Developing Countries*, Washington: National Academies Press, 2005.

López, Luis A. 2007. *Uniones Conyugales y Distancia Social en América Latina. Elementos para su Comprensión en un contexto de Cambio.* Master's Thesis. Centro de Estudios Demográficos, Universidad Autónoma de Barcelona, 2007.

McCaa, Robert. 2000. “Familia y Género en México. 2000 Crítica Metodológica y Desafío Investigativo para el Fin del Milenio,” in Victor Manuel Uribe Urán and Luis Javier Ortiz Mesa (eds.), *Naciones, Gentes y Territorios: Ensayos de Historia e Historiografía Comparada de América Latina y el Caribe*. Medellín: Editorial Universidad de Antioquia, 2000, p. 103-138.

McCaa, Robert, Albert Esteve and Clara Cortina. 2006. “Marriage Patterns in Historical Perspective: Gender and Ethnicity, “ in Ueda Reed (ed.), *A Companion to American Immigration.* Malden, MA: Blackwell Publishers, 2006, p. 359-372.

McCaa, Robert, Steven Ruggles, Michael Davern, Tami Swenson and Krishna Mohan Palipudi. 2006. “IPUMS-International High Precision Population Census Microdata Samples: Balancing

the Privacy-Quality Tradeoff by Means of Restricted Access Extracts," in *Privacy in Statistical Databases*, New York: Springer. 2006, p. 375-382.

McCaa, Robert and Albert Esteve. 2007. "El proyecto IPUMS-International: Microdatos Censales Para Investigadores Argentinos, Latinoamericanos y del Resto del Mundo," in M. Boleda and M. C. Mercado Herrera (eds.), *Seminario Internacional de Población y Sociedad en América Latina, 2005 (SEPOSAL 2005),* 2 Tomos, Salta, Argentina: Tomo I, 2007, p. 51-74.

McCaa, Robert, Albert Esteve and Clara Cortina. 2007. "Gender and Ethnicity: Marriage Patterns in Historical Perspective," in M. Boleda and M. C. Mercado Herrera (eds.), *Seminario Internacional de Población y Sociedad en América Latina, 2005 (SEPOSAL 2005),* 2 Tomos, Salta, Argentina: Tomo I, 2007, p. 37-50.

World Bank. 2006. *World Development Report 2007: Development and the Next Generation*, Washington, DC: The World Bank, 2006.

Woubalem, Zewdu. 2006. *Interim Report: Estimates of Excess Adult Deaths due to HIV/AIDS in Kenya*, IR 06-013, March 2006, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria.

3. Dissertations and Theses

Aray Laverde, Patrick. 2008. "The Brain Drain in Latin America," Masters Thesis in Economics, Université Catholique de Louvain, Belgium. In process.

Becker, Jaime. 2008. "Seeing the Forest: Global Gender Inequality," Ph.D. Dissertation, Sociology Department, University of California at Davis. In process.

Biller, Timothy. 2004. "The Impact of Foreign Direct Investment on Mexico's Agricultural Sector and Forests," Honors Thesis, Economics Department, Tufts University, 2004.

Cruces, Guillermo. 2004. "Poverty Dynamics, Fertility and Labour Supply in Argentina, " Ph.D. Dissertation, Economics Department, London School of Economics and Political Science.

Lanza Queiroz, Bernardo. 2005. *Labor Force Participation and Retirement Behavior in Brazil*, Ph.D. Dissertation, University of California, Berkeley, 2005.

Porter, Maria. 2007. "Imbalance in China's Marriage Market and its Effect on Intra-Household Resource Allocation," Chapter 1 in *Empirical Essays on Household Bargaining in Developing Countries* , Ph.D. Dissertation, University of Chicago Department of Economics, June 2007.

Sanchez, Landy L. 2008. "Permeable walls? Socio-Economic Residential Segregation and Labor Markets Outcomes in Mexico City," Ph.D. Dissertation, Department of Sociology, University of Wisconsin-Madison (August 2008 anticipated). In process.

Zhang, Yuanting . 2007. “Changes in marital dissolution patterns among Chinese and Chinese immigrants: An origin-destination analysis.” Ph.D. Dissertation, Bowling Green State University, 2007.

4. Conference presentations

Alexander, Trent. 2003. “Public Use Census Microdata in Social Science and Social Policy Research.” Midwest Conference on Demographics for Policy Analysts. Minneapolis, April 2003.

Bell, Martin. 2003. “Comparing Internal Migration between Countries: Measures, Data Sources and Results.” Population Association of America, Minneapolis, May 2003.

Block, William C., Colin C. Davis and Marcus G. Peterson. 2003. “The Future of the Integrated Public Use Microdata Series: IPUMS International and IPUMS Redesign.” International Association of Social Science Information Service and Technology, Ottawa, May 2003.

Bryant, John. 2007. “Independent Child Migrants: Some Basic Information and How to Find out More.” UNICEF-University of Sussex Workshop on Independent Child Migrants: Policy Debates and Dilemmas, September 12, 2007, London.

Bullington, Matthew and E. Anthony Eff. 2007. “Domestic Mexican Migration: A Gravity Model,” The Academy of Economics and Finance, Jacksonville, Florida, February 13, 2007.

Cabré, Anna and Albert Esteve. 2004. “Marriage Squeeze and Changes in Family Formation: Historical Comparative Evidence from Spain, France, and United States during the Twentieth Century.” Population Association of America, Boston, April 1-3, 2004.

Carter, Susan B. and Richard Sutch. 2003. “Mexican Fertility Transition in the American Mirror,” Economic History Society Annual Conference, London, April 2-4, 2003.

Chauvel, Louis. 2001. “Education and Class Membership Fluctuation by Cohorts in France and the USA (1960-2000).” ISA RC28 Meeting (International Sociological Association, Research Committee on Social Stratification), Mannheim, Germany, April 26-28, 2001.

Chauvel, Louis. 2001. “Educational Growth and Cohort Changes of Social Structure in France and United-States (1968-2000),” EURESCO (European Science Foundation) Conference, Kerkrade, Netherlands, October 6-10.

Davis, Colin. 2004. “Missing Data Allocation in the IPUMS: Minnesota Allocation Techniques and Customizable Tools for Researchers.” International Association of Social Science Information Service and Technology, Madison, May 2004.

Demont, Floriane and Patrick Heuveline. 2008. “The Cambodian family after the Khmer Rouge Genocide: Continuity and Change,” Population Association of America, New Orleans, April 2008.

Dorn, Sherman. 2007. “Comparative Educational Attainment Portraits, 1940-2002.” Society for the History of Childhood and Youth, Norrkoping, Sweden, June 27-30, 2007.

Esteve, Albert. 2001. “Las experiencias de España y Colombia en IPUMS Internacional.” IPUMS-International Workshop, Bogotá, Colombia, March, 2001.

Esteve, Albert. 2003. “Dios los Cría, ¿y Ellos se Juntan? El Efecto de la Educación en la Homogamia Matrimonial en México, 1970-2000.” Sociedad Mexicana de Demografía, Guadalajara, Mexico, December 2003.

Esteve, Albert. 2004. “Homologacion de las Variables: Microdatos y Metadatos. Simposio Latinoamericano de Homologacion y Divulgacion de Microdatos Censales,” Cartagena, January 2004.

Esteve, Albert. 2004. “El rosto de IPUMS-International en la web.” Simposio latinoamericano de homologacion y divulgacion de microdatos censales, Cartagena, January 2004.

Esteve, Albert, A. Torrents and C. Cortina. 2004. “Proyecto IPUMS, Integrated Public Use of Microdata Series: Aplicabilidad a un Estudio Sobre la Emigración Española a Florida entre 1880 y 1920.” Asociación de Demografía Histórica, Granada, Spain, April 2004.

Esteve, Albert. 2004. “Homologacion de las Variables: Microdatos y Metadatos. Simposio Latinoamericano de Homologacion y Divulgacion de Microdatos Censales,” Cartagena, January 2004.

Esteve, Albert and Robert McCaa. 2006. “Educational Homogamy of Mexicans in Mexico and the USA: Gender, Generation, Ethnicity and Educational Attainment.” Population Association of America, Los Angeles, March 30-April 1, 2006.

Esteve, Albert, Robert McCaa and Anna Cabré. 2006. “The IPUMS-Europe Project: Integrating the Region´s Census Microdata.” European Population Conference, Liverpool, UK, June 21-24, 2006.

Fairlie, Robert W. and Christopher Woodruff. 2007. “Mexican-American Entrepreneurship.” All UC Labor Conference, University of California-Davis, September 2007.

Ferguson, B.D., J.A. Restrepo, and J.H. Jones. 2008. “Missing Men: The Direct Mortality Impacts of Firearm Violence in Colombia, 1979-2005,” Population Association of America Annual Meetings, New Orleans, April 17, 2008.

Garenne, Michel, Robert McCaa and Kourtoum Nacro. 2007. “Maternal Mortality in South Africa in 2001: From Census to Epidemiology.” UAPS Conference (Union of African Population Studies), Arusha, Tanzania, December 10-14, 2007.

Garenne, Michel. 2007. “Situations of Fertility Stall in Sub-Saharan Africa.” UAPS Conference (Union of African Population Studies), Arusha, Tanzania, December 10-14, 2007.

Hall, Patricia Kelly. 2000. "Roundtable on International Historical Microdata." Social Science History Association, Pittsburgh, October 2000.

Hall, Patricia Kelly and Catherine Fitch. 2000. "Roundtable on Definitions of Poverty in Census and CPS Data." Social Science History Association, Pittsburgh October 2000.

Hernandez, Elaine M. and John Robert Warren. 2007. "The Effects of Macro- and Individual-Level Socioeconomic Status on Child Mortality in Brazil, 1970 and 2000." International Sociological Association Research Committee on Social Stratification and Mobility RC28. Montréal, Canada. August 2007.

Jones, J.H. and B.D. Ferguson. 2007. "The Consequences of Sex-Ratio Imbalances on Dyadic Power within Colombian Households," American Anthropological Association Annual Meeting, Washington, DC, November 30, 2007.

King, Mary C. 2008. "Mexican Women's Work on Both Sides of the U.S. Mexican Border." Allied Social Science Associations, New Orleans, January 4-6, 2008.

King, Miriam. 2001. "People are the Problem: Human Demographic and Economic Data Sources for Ecological Research." Ecological Society of America, Madison, Wisconsin, August 2001.

Lagakos, David. 2007. "Market Size, Productivity, and Technology Adoption: The Case of the Retail Sector." University of California-Santa Barbara Conference on Latin American Productivity, Santa Barbara, September 2007.

Lam, David and Letícia Marteleto. 2008. "Family Size of Children and Women during the Demographic Transition," Population Association of America, New Orleans, April 16-19 2008.

Lam, David and Letícia Marteleto. 2005. "Stages of the Demographic Transition from a Child's Perspective: Family Size, Cohort Size, and Children's Resources." IUSSP International Population Conference (International Union for the Scientific Study in Population), Tours, France, July 2005.

Lambert, Paul S., Kenneth Prandy and Manfred Max Bergman. 2005. "Specificity and Universality in Occupation-Based Social Classification." European Association for Survey Research, Barcelona, July 18-22, 2005.

Martínez Gómez, Ciro L. 2000. "El Uso de los Microdatos Censales. Una Aplicación a la Migración Interna en Colombia," Simposio de Estadistica 2000: Censos, Encuestas y Sistemas de Información Estadística, San Andrés, Colombia, August 5-12, 2000.

Martínez Gómez, Ciro L. 2001. "Variables De Clasificación Geográfica," Taller col-IPUMS, Homologación de los datos censales de Colombia, Bogotá, Colombia, March 23-24.

McCaa, Robert and Steven Ruggles. 2000. "A Reality Check for IPUMS: Labor Force Participation of Mexican Women in Mexico - Census Microdata versus Employment Survey." CCSR Center for Social Science Research, "The Census of Population: 2000 and Beyond," Manchester, UK, June 22-23, 2000.

McCaa, Robert. 2000. “IPUMS-International: A Report on the First Year.” Workshop on The National Censuses: An International Research Tool? How Can We Achieve Comparability? World History Congress, Oslo, Norway, August 2000.

McCaa, Robert and Steven Ruggles. 2001. “The Census in Global Perspective and the Coming Microdata Revolution.” The 14th Nordic Demography Symposium, Tjøme, Norway, May 2001.

McCaa, Robert, Rodolfo Gutiérrez and Gabriela Vásquez. 2001. “Women in the Workforce: Calibrating Census Microdata against a Gold Standard: Mexico 1990 and 2000.” International Union for the Scientific Study of Population World Conference, Bahia, Brazil, August 2001.

McCaa, Robert, Steven Ruggles and Matthew Sobek. 2002. “Using Census Microdata: The IPUMS International Project.” Association of National Census and Statistics Directors of America, Asia, and the Pacific, Ulaanbaatar, Mongolia, June 2002.

McCaa, Robert and Nikolai Botev. 2003, “Integrating European Census Microdata.” Working Party on Demographic Statistics and Population and Housing Censuses, Eurostat. Luxembourg, February 2003.

McCaa, Robert, Murungaru Kimani, Albert Esteve, Jose Rodolfo Gutierrez-Montes and Gabriela Vazquez-Benitez. 2003. “Calibrating Census Microdata Against Gold Standard Surveys: Kenya 1999 (Fertility) and Mexico 2000 (Female Labor Force).” Population Association of America, Minneapolis, May 2003.

McCaa, Robert, Steven Ruggles, Matt Sobek and Albert Esteve. 2003. “IPUMS-International: A Restricted Access Web-Site Providing Anonymized, Integrated Census Microdata for Social Science and Policy Research.” International Statistical Institute, Berlin, August, 2003.

McCaa, Robert and Albert Esteve. 2003. “El proyecto IPUMS-International: Microdatos censales para investigadores y planificadores en Chile, Latino América y el mundo.” Seminario Internacional IASI 'Estadistica y Desarrollo Local en un Mundo Globalizado, Valdivia, Chile, October 2003.

McCaa, Robert, Steven Ruggles, Matt Sobek and Albert Esteve. 2003. “IPUMS-Asia/Pacific: Synopsis of a Proposal,” ANCSDAAP (Association of National Census and Statistics Directors of America, Asia and the Pacific), 21st Population Census Conference: Analysis of the 2000 Round of Censuses, Kyoto, Japan, November 19-21, 2003.

McCaa, Robert. 2003. “Family relationships in Mexican Censuses: A Proposal for the International Integration of Census Microdata and a Historical Overview.” Sociedad Mexicana de Demografía, Guadalajara, Mexico, December 2003.

McCaa, Robert. 2004. “Using IPUMS-International: A Restricted Access Web-Site Offering Anonymized, Integrated Census Microdata of China, the United States, Mexico, Brazil, France, and Other Countries Free of Charge.” International Institute of Sociology World Congress, Beijing, July 7-11, 2004.

McCaa, Robert and Rodolfo Gutierrez. 2005. "Harmonized Census Microdata of Mexico and the USA: A Comparison of Women in the Workforce by Birthplace, Origin and Ethnicity." American Historical Association Annual Meeting, Seattle, January 2005.

McCaa, Robert and Agnes Odinga. 2005. "Kenyan Census Microdata: Orphanhood as an Illustration of Research Opportunities and Challenges." American Historical Association Annual Meeting, Seattle, January 2005.

McCaa, Robert, Steven Ruggles and Matthew Sobek. 2005. "IPUMS-International Harmonized Census Microdata Extract System: Users and Uses, May 2002-January 2005." ANCSDAAP 2005 Conference, Seattle, March 2005.

McCaa, Robert, Albert Esteve and Clara Cortina. 2005. "Gender and Ethnicity: Marriage Patterns in Historical Perspective." Seminario International de Población y Sociedad, Salta, Argentina, June 2005.

McCaa, Robert and Agnes Odinga. 2005. "Statistical Confidentiality and the Dissemination of Restricted-Access Integrated Census Microdata Extracts: The Case of Kenya, 1969-1999." International Commission for Historical Demography, Sydney, Australia, July 2005.

McCaa, Robert and Albert Esteve. 2005. "Homogamia Educacional en México y Brasil, 1970-2000: Pautas y Tendencias." International Union for the Scientific Study of Population, XXV International Population Conference, Tours, France, July 2005.

McCaa, Robert, Felicien Donat E. T. Accrombessy and Khassoum Diallo. 2005. "Calibrating Orphanhood: The Number of Orphans According to Recent Censuses and Health Surveys already Exceed UNAIDS Estimates for 1020 for Kenya and Benin and 4/5th for South Africa," Global Forum for Health Research IX, Mumbai, India, September 12-16, 2005.

McCaa, Robert, Steven Ruggles and Matthew Sobek. 2005. "IPUMS-International: Making Confidentialized, Harmonized Census Microdata for 44 Countries Available Free-of-Charge to Academic and Policy Researchers World-Wide." Fourteenth Conference of Commonwealth Statisticians, Capetown, South Africa, September 2005.

McCaa, Robert, Steven Ruggles, Michael Davern, Tami Swenson and 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 2006 (PSD"2006), Rome, Italy, December 13-15, 2006

McCaa, Robert. 2006. "IPUMS: la Familia Crece." 2º Seminario Internacional: Efectos de la Globalización y las Políticas Migratorias, Toluca, Mexico, November 15-17, 2006.

McCaa, Robert. 2006. "Disseminating Integrated, Anonymized, High Precision Census Microdata Samples: an Invitation, Update and Proposal." Fourteenth Meeting of the Regional Census Coordinating Committee (CARICOM), Port-of-Spain, Trinidad and Tobago, November 9-10, 2006.

McCaa, Robert. 2006. "Indigenous Peoples, Ethnicity and Identities in Contemporary Censuses: A Global Perspective." Indigenous Identities in Demographical Sources, Umeå, Sweden, September 29-30, 2006.

McCaa, Robert and Albert Esteve. 2006. "Homogamia Educativa de los Mexicanos en México y Estados Unidos: Género, Generación, Origen y Educación." Reunión Anual de la Sociedad Mexicana de Demografía (SOMEDE), Guadalajara, Mexico, September 5-9, 2006.

McCaa, Robert and Albert Esteve. 2006. "Homogamia Educacional en México y Brasil, 1970 – 2000: Pautas y Tendencias." II Congreso de la Asociación Latinoamericana de Población, Guadalajara, Mexico, September 3-5, 2006.

McCaa, Robert. 2006. "Disseminating Integrated Census Microdata to Academic Researchers and Policy Makers at No Cost." Workshop on Advocacy and Resource Mobilization for Phase I (2005-2009) of the 2010 Round of Population and Housing Censuses in Asia, Phnom Penh, Cambodia, July 25–28, 2006.

McCaa, Robert, Albert Esteve, Steven Ruggles, Matt Sobek and Ragui Assaad. 2006. "Using Integrated Census Microdata for Evidence-based Policy Making: the IPUMS-International Global Initiative." Indian Association for Social Sciences and Health, Third All India Conference, New Delhi, March 16-18, 2006.

McCaa, Robert, Steven Ruggles and Matt Sobek. 2006. "Disseminating Census Microdata: an Essential Component of National Strategies for the Development of Statistics." Forum on African Statistics Development (FASDEV-II), Addis Ababa, February 6-10, 2006.

McCaa, Robert, Steven Ruggles and Matt Sobek. 2006. "Archiving Census Microdata: The IPUMS-International Strategy." Forum on African Statistics Development (FASDEV-II), Addis Ababa, February 6-10, 2006.

McCaa, Robert, Steven Ruggles and Matt Sobek. 2007. "IPUMS-International Integrated Census Microdata Extract System: Users and Uses." May 2002-March 2007." 23rd ANCSDAAP Population Census Conference, Christchurch, New Zealand, April 16-18, 2007.

McCaa, Robert, Steven Ruggles and Matt Sobek. 2007. "Using Census Microdata Disseminated by IPUMS-International to Assess Millennium Development Goals of Literacy, Education and Gender Equity in the Ugandan censuses of 1991 and 2002." Scientific Statistics Conference, Kampala, Uganda, June 11-13, 2007.

McCaa, Robert and Krishna Mohan Palipudi. 2007. "Integrating Disability Census Microdata: What is Accessible from IPUMS-International?"56th Session of the International Statistical Institute, Lisbon, Portugal, Aug. 22-29, 2007.

McCaa, Robert. 2007. "IPUMS-International High Precision Population Census Microdata Samples: Balancing the Privacy-Quality Tradeoff by Means of Restricted Access Extracts." International Conference on Quality Management of Official Statistics, Daejeon, Republic of Korea, Sep. 6-7, 2007.

McCaa, Robert and Antonio López Gay. 2007. "Género y trabajo en los censos de población de América Latina: la captación de la actividad económica femenina secundaria a partir de la ampliación del cuestionario censal con una única pregunta," UNFPA Workshop on Preparatory Activities, Analysis and Exchange of Experiences for the Successful Implementation of the 2010 Round of Population and Housing Censuses in Latin America and the Caribbean, Panama City, Panama, Sept. 17-21.

McCaa, Robert and Awad Hag Ali. 2007. "Preserving Census Microdata and Making Them Useful: Sudan." Arab Statistical Conference, Amman, Jordan, November 12-13, 2007.

Minca, Elisabeta. 2008. "Patterns of Romanian Emigration: Is there a Brain Drain?" European Population Conference, Barcelona, July 9-12, 2008.

Miranda-Ribeiro, Adriana, E. Rios-Neto and J.A. Ortega. 2006. "Declíno da Fecundidade no Brasil e México e o Nível de Reposição: Efeito Tempo, Quantum e Parturição." II Congreso de Asociación Latino Americana de Población, Guadalajara, Mexico, September 3-5, 2006.

Mueller, Richard E. and Gary L. Hunt. 2003. "Canadian Immigration to the U.S., 1985-1990: Estimates from a Roy Selection Model of Differences in Returns to Skill," with Gary L. Hunt, Conference on Macro Policy, Koç University and Sabanc University, Istanbul, Turkey, August 18-19, 2003.

Mueller, Richard E. 2005. "What Happened to the Canada U.S. Brain Drain of the 1990s? New Evidence from the 2000 U.S. Census," Nordic Association for Canadian Studies, 8th Triennial Conference, Turku, Finland, August 18-20, 2005.

Mueller, Richard E. 2005. "The Migration of Highly Skilled Individuals Within and Between Canada and the United States," with Gary L. Hunt, 21st Century COE Program, Kobe University and Japanese Economic Policy Association Joint International Conference, Awaji Island, Japan, December 17-18, 2005.

Mueller, Richard E. 2006. "The Migration of Highly Skilled Individuals Within and Between Canada and the United States," with Gary L. Hunt, 8th National Metropolis Conference, Vancouver, British Columbia, March 23-26, 2006.

Mueller, Richard E. 2006. "The Migration of Highly Skilled Individuals Within and Between Canada and the United States," with Gary Hunt, International Mobility of Highly Skilled Workers: HRSDC-IC-SSHRC, Skills Research Initiative Workshop, Ottawa, Ontario, June 9, 2006.

Mueller, Richard E. 2006. "A Note on Canadian Migration to the United States During the 1980s and 1990s," All China Economics International Conference, Hong Kong, China, December 18-20, 2006.

Odinga, Agnes and Robert McCaa. 2001. "Statistical Confidentiality and the Construction of Anonymized Public Use Census Samples: a Draft Proposal for the Kenyan Microdata for 1989." Social Science History Association, Chicago, November 2001.

Peterson, Marcus. 2004. "Getting Wired: Caffeinating Microdata Production at the Minnesota Population Center with Java." International Association of Social Science Information Service and Technology, Madison, May 2004.

Prandy, Ken, Paul S. Lambert and Manfred Max Bergman. 2002. "National Contexts and Cross-National Comparisons of Structures of Social Stratification." XVth ISA World Congress of Sociology, Research Committee 20 (Comparative Research), Brisbane, Australia, July 12, 2002.

Rios-Neto, E. and Adriana Miranda-Ribeiro. 2007. "Fertility Decline in Brazil and Mexico: Tempo, Quantum and Parity Composition Effects." Population Association of America, New York, NY, March 29-31, 2007.

Ruggles, Steven. 2004. "The Microdata Revolution: A brief History." Simposio Latinoamericano de Homologacion y Divulgacion de Microdatos Censalesm, Cartagena, January 2004.

Ruggles, Steven and Misty Heggeness. 2008. "Intergenerational Families in Developing Countries," Population Association of America, New Orleans, April 16-19 2008.

Ruggles, Steven. 2008. "Living Arrangements of the Aged in Comparative Historical Perspective," European Social Science History Conference, Lisbon, February 27, 2008.

Ruggles, Steven. 2007. "Intergenerational Coresidence in Developing Countries: A Comparative Historical Perspective," Social Science History Association, Chicago, November 15-19, 2007.

Ruggles, Steven. 2007. "The Relationship of Socioeconomic Status to Intergenerational Coresidence: A Comparative Historical Analysis." PAA Population Association of America, New York, March 29-31, 2007.

Ruggles, Steven. 2007. "Using Cyberinfrastructure to Develop Databases for Social Science Research," American Association for the Advancement of Science, San Francisco, February 16-19 2007.

Ruggles, Steven. 2006. "The Case for Open Access to Data," Presented at "Disseminating and Analyzing Longitudinal Historical Data," International Institute for Social History, Amsterdam, March 21, 2006.

Ruggles, Steven and Catherine Fitch. 2000. "International Integrated Microdata Access System." IASSIST 2000: Data in the Digital Library, Evanston, Illinois, June 2000.

Salvia A. and P. De Grande P. 2007. "Segregación Residencial Socioeconómica y Espacio Social: Deserción Escolar de los Jóvenes en el Área Metropolitana del Gran Buenos Aires." XXVI Congreso Asociación Latinoamericana de Sociología, Guadalajara, México, August 12-18, 2007.

Sobek, Matthew. 2004. "Dataset Processing: Standardizing the Input: Analysis, Reformatting, Drawing Samples." Simposio Latinoamericano de Homologacion y Divulgacion de Microdatos Censales, Cartagena, January 2004.

Sobek, Matthew, Trent Alexander and Carolyn Liebler. 2003. "Using the Integrated Public Use Microdata Series (IPUMS) in Research." American Sociological Association, Atlanta, August 2003.

Sobek, Matthew, Robert McCaa and Albert Esteve. 2002. "The IPUMS-International Project: Challenges and Methods of International Census Data Integration." Social Science History Association, St. Louis, October 2002.

Thomas, Wendy L. and Robert McCaa. 2001. "Archiving Census Documentation and Microdata: Preserving Memory, Increasing Stakeholders." United Nations Symposium on Global Review of 2000 Round of Population and Housing Censuses: Mid-decade Assessment and Future Prospects, New York, August 2001.

Turra, Cassio M. and Bernardo L. Queiroz. 2005. "Before It's Too Late: Demographic Transition, Labor Supply, and Social Security Problems In Brazil." United Nations Expert Group Meeting on Social and Economic Implications of Changing Population Age, Population Division, Department of Economic and Social Affairs, United Nations Secretariat, Mexico City, August 31-September 2, 2005.

5. Workshops and lectures:

Alexander, Trent. 2003. "Historical Census Data, 1850-2000: Accessing and Using Resources." Seminar offered in the Summer Program in Quantitative Methods at the Inter-university Consortium for Social and Political Research (ICPSR), Ann Arbor, June 2003.

Alexander, Trent and Bill Block. 2004. "New Data Projects at the Minnesota Population Center." Social Sciences Faculty Seminar Series, Carleton College, Northfield, Minnesota, June 2004.

Alexander, Trent and Patricia Kelly Hall. 2003. "Using Census Microdata in Social Policy Research." Two-week course offered at the Hubert H. Humphrey Institute of Public Affairs, University of Minnesota. Spring 2003.

Alexander, Trent and Patricia Kelly Hall. 2003. "Using Census Microdata in Social Policy Research." Two-week course offered at the Hubert H. Humphrey Institute of Public Affairs, University of Minnesota. Fall 2003.

Alexander, Trent and Patricia Kelly Hall. 2004. "Using Census Microdata in Social Policy Research." Two-week course offered at the Hubert H. Humphrey Institute of Public Affairs, University of Minnesota. Spring 2004.

Alexander, Trent and Evan Roberts. 2004. "Work, Family, and Community: Global Perspectives in Examining Population History." Teacher Summer Institute course, Institute for Global Studies, University of Minnesota, July 2004.

Chauvel, Louis. 2003. "Génération Sociale et Socialisation Transitionnelle: Fluctuations Cohortales et Stratification Sociale en France et aux Etats-Unis au XXe siècle." Institut d"Etudes Politique de Paris, 2003.

Esteve, Albert. 2004. "Marital Homogamy in Mexico: The Impact of Education, 1970-2000." Brown Bag Seminar series, Department of Demography, University of California at Berkeley, January 2004.

Esteve, Albert, Robert McCaa, Steven Ruggles and Matt Sobek. 2005. "International Comparisons Based on Census Microdata (IPUMS): Methods and Applications." Les Lundis de l"INED seminar presentation, June 6, 2005, Paris.

Golaz, Valerie. 2007. "Presentation of IPUMS-International and the Use of Kenyan Data Bases." Session in workshop on quantitative methods for Ph.D. students from East African Universities, sponsored by L"institut français de recherche en Afrique and (IFRA) "Institut de recherche pour le développement (IRD), Nairobi, Kenya, July 2007.

McCaa, Robert. 2002. "Women in the Workforce: Calibrating Census Microdata and Employment Surveys: Mexico 1970, 1990 and 2000." Lecture, Population Studies Center, University of Michigan, Ann Arbor, December 2002.

McCaa, Robert. 2003. "Historia y Demografia: Reflecciones y Lecciones del Proyecto IPUMS-Internacional, el Caso de Mexico." Dialogos con el Pensamiento Historiador Colloquim, Universidad Autonoma de Puebla, Puebla, Mexico, June 2003.

McCaa, Robert. 2004. "Women in the Workforce: Calibrating Census Microdata against Gold Standards: Mexico 1990-2000." Bureau of Labor Statistics, January 2004.

McCaa, Robert. 2005. "IPUMS-International: Project Goals and How We Accomplish Them." IPUMS-International Asian and Pacific Workshop, Seattle, March 2005.

Ruggles, Steven. 2007. "International Censuses and Intergenerational Families," Plenary Address, Human and Social Dynamics Conference, National Science Foundation. Arlington, VA, October 1-2 2007.

Sobek, Matthew. 2005."Teaching with the Integrated Public Use Microdata Series." Midwest Sociological Society Meetings, Minneapolis, April 2005.

Sobek, Matthew. 2005. "Research Workshop: Using the Integrated Public Use Microdata Series in Research." American Sociological Association, Philadelphia, August 2005.

APPENDIX E
IPUMS-International Data Inventory

1. Summary table

Status	Number of countries	Number of samples
A. Fully processed	26	80
B. May 2008 release	14	32
C. Other samples received	45	98

2. List of samples

Note: Samples in **Bold** have shifted categories since the last report

A. Fully processed: 80 samples from 26 countries

Argentina	**1970, 1980, 1991, 2001**
Belarus	1999
Brazil	1970, 1980, 1991, 2000
Cambodia	1998
Chile	1960, 1970, 1982, 1992, 2002
China	1982
Colombia	1964, 1973, 1985, 1993
Costa Rica	1963, 1973, 1984, 2000
Ecuador	1962, 1974, 1982, 1990, 2001
France	1962, 1968, 1975, 1982, 1990
Greece	1971, 1981, 1991, 2001
Hungary	**1970, 1980, 1990, 2000**
Israel	**1972, 1983, 1995**
Kenya	1989, 1999
Mexico	1960, 1970, 1990, 2000
Palestine	**1997**
Philippines	1990, 1995, 2000
Portugal	**1981, 1991, 2001**
Romania	1992, 2002
Rwanda	**1991, 2002**
South Africa	1996, 2001
Spain	1981, 1991, 2001
Uganda	1991, 2002
United States	1960, 1970, 1980, 1990, 2000
Venezuela	1971, 1981, 1990
Vietnam	1989, 1999

B. Scheduled for release May 2007: 32 samples from 14 countries (9 net)

Austria	**1971, 1981, 1991, 2001**
Canada	**1971, 1981, 1991, 2001**
China	**1982, 1990**
Colombia	**2005**
Egypt	**1996**
Ghana	**2000**
Iraq	**1997**
Malaysia	**1970, 1980, 1991, 2000**
Mexico	**1995, 2005**
Netherlands	**1960, 1970, 2001**
Panama	**1960, 1970, 1980, 1990, 2000**
United Kingdom	**1991, 2001**
United States	**2005**
Venezuela	**2001**

C. Additional data received by MPC: 98 samples from 45 countries (37 net)

Armenia	2000
Bangladesh	**1991**
Bolivia	1976, 1992, 2001
Botswana	**2001**
Czech Republic	1991, 2001
Dominican Rep.	1960, 1970, 1981
El Salvador	1992
Egypt	1986
Ethiopia	**1994, 1984**
Fiji	1966, 1986, 1996
France	1999
Germany	**1971**
Ghana	**1984**
Guinea	1983, 1996
Guatemala	1973, 1981
Haiti	**1971, 1982, 2003**
Honduras	1961, 1974, 1988, **2001**
India	**1983, 1987, 1993, 1999, 2005**
Indonesia	1971, 1976, 1980, 1985, 1990, 1995, 2000
Israel	1961
Italy	1981, 1991
Kenya	1979
Liberia	**1974**
Madagascar	1993
Malawi	1987, 1998
Mali	**1976**, 1987, 1998
Mauritius	1990, 2000
Mexico	1980

C. Additional data received by MPC: 98 samples from 45 countries (continued)

Mongolia	**1989**, 2002
Nepal	**2001**
Nicaragua	1971
Pakistan	1973, 1981, 1998
Paraguay	1962, 1972, 1982, 1992, 2002
Peru	1993
Philippines	1960, 1970, 1980
Puerto Rico	1970, 1980, 1990, 2000
Romania	1977
Saint Lucia	1982, 1991
Sierra Leone	**2004**
Slovenia	**2002**
Sudan	1973, 1983, 1993
Tanzania	**1988**
Thailand	1970, 1980, **1985**, 1990, 2000
Turkmenistan	1995
Uruguay	1963, 1975, 1985, 1996

IPUMS-International Data Producers Workshop
New York City (Columbia University)
February 23, 2008

AITRS = Arab Institute for Training in Research and Statistics
ISI = International Statistics Institute
CIESIN = Center for International Earth Science Information Network
SIPA = School of International and Public Affairs
ISERP = Institute for Social and Economic Research and Policy
USCB = US Census Bureau

1	Ben	Kiregyera	Director, African Center for Statistics (UNECA)
2	Dr. Hilal	Al-Bayyati	AITRS
3	Ibtisam I.	Khalil	AITRS
4	Susan	Linacre	Australia
5	Gillian	Nicoll	Australia
6	Anna	Majelantle	Botswanna
7	AYM Ekramul	Hoque	Bangladesh
8	Eduardo	Pereira	Brazil
9	Zelia	Bianchi	Brazil
10	Jane	Badets	Canada - StatsCan
11	Dirk	Jaspers	CELADE
12	Héctor	Maldonado	Colombia
13	Santiago	Molina	Colombia
14	John	Coatsworth	Columbia
15	Susana	Adamo	Columbia -CIESIN
16	Dominique	Kimpouni	Congo
17	Jaroslav	Kraus	Czech Republic
18	Dennis	Trewin	Dennis Trewin Statistical Consulting
19	Dario Antonio	Lopez Villar	Dominican Republic
20	Khaled Ahmed	El Said	Egypt - CAPMAS
21	Samia Zekaria	Gutu	Ethiopia
22	Timoci	Bainimarama	Fiji Islands
23	Andrea	Harusz	Germany
24	Grace	Bediako	Ghana
25	Nicholas	Nsowah-Nuamah	Ghana
26	Marie Ann	Doualamou	Guineé
27	Kenneth	Hill	Harvard
28	Michael	Levin	Harvard/USCB
29	Wai Kong	Tang	Hong Kong, China
30	Evelio	Fabbroni	IASI
31	Dr. S. K.	Nath	India
32	Rusman	Heriawan	Indonesia
33	Mr. Wynandin	Imawan	Indonesia
34	Denise	Lievesley	President, ISI
35	Annette	McKenzie	Jamaica
36	Bunzo	Hirai	Japan
37	Wajdi	Ibraim	Jordan
38	Dr. Gazi	Shbaikat	Jordan
39	Myung Jin	Hwang	Korea
40	Nurbek	Tulegabylov	Kyrgyz Republic

IPUMS-International Data Producers Workshop
New York City (Columbia University)
February 23, 2008

41	Ms. Aishath	Shahuda	Maldives
42	Fuwad	Thowfeek	Maldives
43	Chris	Muller	Muller Media
44	Francis	Vierbergen	Netherland Antilles
45	Mr. Geert	Bruinooge	Netherlands
46	Ms. Ada	van Krimpen	Netherlands
47	Geoff	Bascand	New Zealand
48	Dr. Luay	Shabaneh	Palestine
49	Jozef	Olenski	Poland
50	Vergil	Voineagu	Romania
51	Pali	Lehohla	South Africa
52	John	Male Mukasa	UBOS - Uganda
53	Kirill	Andreev	UN Population Division
54	Patrick	Gerland	UN Population Division
55	Taeke	Gjaltema	UN Population Division
56	Jose Antonio	Ortega	UN Population Division
57	François	Pelletier	UN Population Division
58	Cheryl	Sawyer	UN Population Division
59	David	Glejberman	Uruguay
60	Laura	Zayatz	USCB-Disclosure Review
61	Olivier	Dupriez	World Bank Development Group
62	Efreda	Chulu	Zambia