IPUMS-International Response to Advisory Board

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

September 2007

Minnesota Population Center

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.

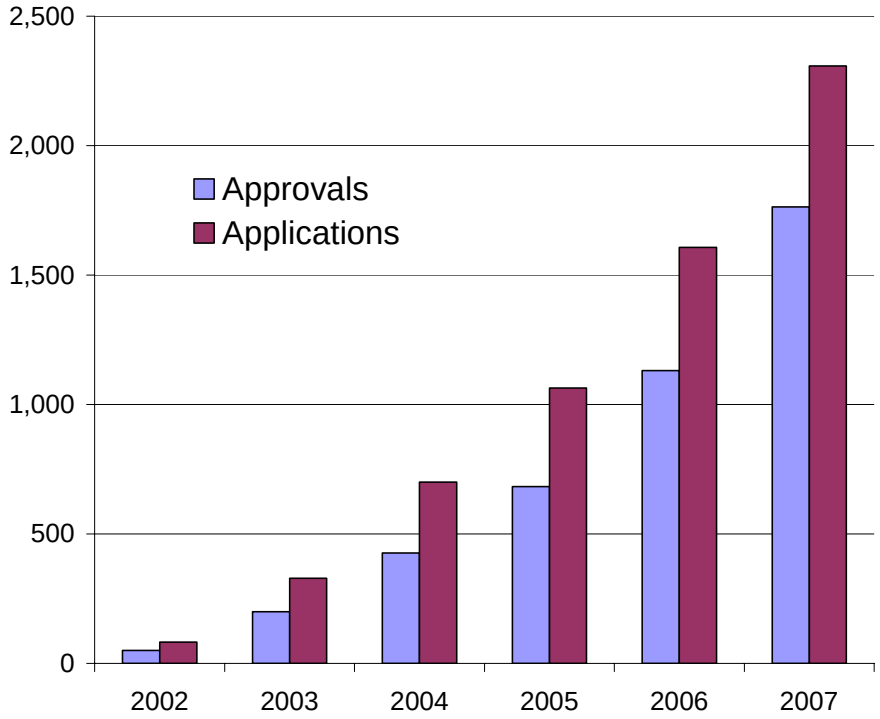
We have enlisted the assistance of data producers to reach new users. We have held multi-day data producer workshops in Seattle (March 2005), Paris (June 2006), and Lisbon (August 2007). While the primary purpose of these workshops is to communicate with our partners and obtain new data and data dissemination 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 2005, 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].

Appendix:
Publications and Presentations Based on IPUMS-International

1. Journal publications 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.

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. 2007. "Gendered Selectivity: U.S. Mexican Immigrants and Mexican Non-Migrants, 1960-2000," submitted to *Latin American Research Review*, 2007.

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.

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.

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.

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.

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

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

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

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.

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

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.

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.

2. Books and other one-time publications

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

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.

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

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.

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.

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

3. 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.

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.

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.

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.

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. 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.

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

Marínez 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.

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, Duadalajara, Mexico, September 3-5, 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. 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 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.

4. 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.

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.