IPUMS-International Progress Report and Work Plan

Submitted to the IPUMS-International Advisory Board
March 13, 2006

Minnesota Population Center
University of Minnesota

CONTENTS

Acknowledgements. This report was prepared by Steven Ruggles, Matthew Sobek, and Monty Hindman with the assistance of Catherine Fitch, Miriam King, Deborah Levison, and Robert McCaa. The project is funded by NSF grant 0433564 and NIH grants R01 HD047283 and HD044154.

Background

This document describes the procedures and software the IPUMS-International team has developed during the past year for processing census microdata, and outlines our production schedule for the next five years. We also report our progress and plans for other key components of the project, including acquisition, preservation, and dissemination of census microdata.

The project began in 1999 with a social science infrastructure grant from the National Science Foundation, "Integrated International Microdata Access System" (9908380). Our goal was to show the feasibility of preserving the world's census microdata resources and democratizing access to these resources. The project created a comprehensive inventory of known microdata—described in our award-winning *Handbook of International Historical Microdata*—and preserved microdata from over 100 censuses. To demonstrate the potential for international census integration, we selected eight countries with broad geographic dispersion: Brazil, China, Colombia, France, Kenya, Mexico, the United States, and Vietnam. Developing anonymized microdata files suitable for public use involved standardizing formats and correcting format errors, drawing samples, correcting inconsistent and missing responses, assessing confidentially risks and applying protections, and harmonizing coding across countries and censuses. Between 2003 and 2005, IPUMS-International released 28 census samples from these countries, providing data on 156 million persons. These data, together with the accompanying documentation covering temporal and international comparability issues, are freely available to researchers through a new web-based data access system (http://ipums.org/international).

IPUMS-International is already an important data resource. Over 750 projects by scholars in over 40 countries are underway. In addition to university-based researchers, the IPUMS-International user list includes representatives of many national statistical offices and international agencies such as the World Health Organization, the International Labour Office, and the World Bank. Research topics include the changing living arrangements of the aged, female labor-force participation and educational attainment, regional inequality differentials, patterns of age hypergamy, international migration, effects of emigration on labor markets, and relationships between divorce and family composition, between disease factors and education, and between educational attainment and cohort size. Most of these studies incorporate both cross-national and cross-temporal comparisons. For example, a National Academy of Sciences book, *Growing Up Global: Transitions to Adulthood in Developing Countries*, used IPUMS-International data from Colombia, Kenya, Mexico, and Vietnam to analyze changing outcomes such as attending school, working, childbearing, and marrying as a function of age, gender, and household characteristics (Lloyd 2005).

In 2003 and 2004, we received substantial new funding to expand the database dramatically.[1] Already, we have assembled one of the largest collections of population microdata in the world. We have obtained and archived data covering over a billion person

[1] "International Integrated Microdata Series," NSF 0433654; "Integrated Samples of European Censuses," NIH R01 HD047283; "Integrated Samples of Latin American Censuses, 1960-2003," NIH R01 HD044154.

records from 140 censuses in 50 countries, and the datasets are still flowing in. The University of Minnesota has a perpetual license agreement with each country, allowing the Minnesota Population Center (MPC) to redistribute the data for research and educational purposes.

The dramatic increase in the scale of the project has demanded development of streamlined metadata, software, and data processing protocols. It took us almost a year to design and implement the software and procedures that have allowed us to ramp up production. The new systems are now in place, and data processing is proceeding rapidly. Over the next five years, MPC will release microdata samples for dozens of countries around the world, quintupling the world's total quantity of public use microdata, democratizing access to these vital scientific resources, and creating unprecedented opportunities for global comparative research. This document describes how we plan to get the job done.

We begin by describing the software and metadata infrastructure that underlies all aspects of IPUMS-International data processing and dissemination. We then summarize the work process we developed during the past year and explain our proposed plan for data production over the next four years. Finally, we report on the acquisition, preservation, and use of census microdata and on our outreach to researchers.

Software and Metadata Infrastructure

This project has required us to develop a substantial body of new software and metadata. Design of our software systems was carried out under the direction of Bill Block, Peter Clark, Monty Hindman, Catherine Ruggles, and Matt Sobek. On August 17, 2005, Professor Jeffrey Naughton of the University of Wisconsin's Department of Computer Science made a site visit to MPC to evaluate our efforts and offer advice; his thoughtful report is included as Appendix C. Professor Jaideep Srivastava of the University of Minnesota Department of Computer Science has also taken a great interest in the software challenges posed by the project. Professor Srivastava has reviewed our approach, and we have hired one of his graduate students, Nupur Bhatnagar, to work on the project. We look forward to continued collaboration with Professor Srivastava on the design and implementation of a new data model.

IPUMS-International software can be grouped into four principal categories:

1. 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 within the past year.

2. 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. Most of these programs are also new this year.

3. 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 have revised this software substantially during the past year 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.

4. 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. During the past year, 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.

All the software for data preparation, data conversion, and dissemination is driven by metadata. Metadata is formally structured documentation of digital data. During the past year, we have developed a comprehensive metadata system for IPUMS-International. The goal of the system is to capture 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 will be able to 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:

1. 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.

2. Variable Translation Tables. This metadata component provides most of the variable-level information required to create the database, including IPUMS-format variable labels,

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

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.

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

4. 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.

5. Ancillary documentation. This component provides information on enumeration instructions, sample designs, and other material related to the particular census or sample.

To give a sense of what the metadata looks like, Figure 1 shows a snippet of one of the source data dictionaries for the marital status variable in the 2000 census of Costa Rica. Each element in the XML document—the variable name, the variable label, the variable label in the original language, the column location, and so on—is wrapped within a set of tags. Each set of tags is identified by brackets; for example the variable name is identified within the tags <var> and </var>. Moreover, the tags themselves are hierarchically organized in a logical structure. The tags are nested, so that, for example, the variable to which a specific value label refers can be inferred from its position. With relatively little effort programmers can draw in this information, capitalizing on XML functionality built into modern programming languages. The system is flexible, so that new fields can be added or files can be reorganized with minimal difficulty.

Because the XML tags have a defined structure, one can write validation routines to ensure that metadata is properly structured, all expected elements are present, and keys between the different file types match. The logical organization of the XML structure also ensures that informational items are stored in only one place in the system. Both the web and data conversion systems read the same metadata, understand its structure, and pull data out of the single place where each item is stored.

This system is ideal for machine processing, but it is clumsy for humans to edit or read. Although XML offers great advantages for software development and database management, the tags create a need for specialized metadata-creation software. Research staff do not enter XML tags manually because a heavily-tagged document is hard to navigate and edit and because they may accidentally introduce errors into the highly-structured document.

For viewing and editing by humans, we display the information in tabular format without tags. We are using Microsoft Excel and Word as the primary tools to display and maintain most IPUMS-International metadata. Figure 2 shows the same marital status information that appears in Figure 1—together with additional variables relating to employment—without tags, in the form of an Excel spreadsheet. This format makes it easy for the research staff to create, view, and maintain the metadata. To convert between Excel or Word and the tagged XML version of the metadata, we have built a library of VBA macros. Some macros apply tags to documents as they are exported from Excel or Word into XML

Figure 1. Costa Rica 2000 Source Data Dictionary XML (Marital Status part)

```
<variable>
    <var>marst</var>
    <lab>Marital Status</lab>
    <labor>P13-Estado Conyugal</labor>
    <recordtype>P</recordtype>
    <col>111</col>
    <wid>1</wid>
    <frm></frm>
    <svar>CR00A420</svar>
    <sel>0</sel>
    <notes></notes>
    <freqv></freqv>
    <row>2504</row>
    <value>
        <val>1</val>
        <lab>Consensual union</lab>
        <labor>Unido(a)</labor>
        <freq>38937</freq>
    </value>
    <value>
        <val>2</val>
        <lab>Married</lab>
        <labor>Casado(a)</labor>
        <freq>108617</freq>
    </value>
    <value>
        <val>3</val>
        <lab>Separated</lab>
        <labor>Separado(a)</labor>
        <freq>8975</freq>
    </value>
    <value>
        <val>4</val>
        <lab>Divorced</lab>
        <labor>Divorciado(a)</labor>
        <freq>6264</freq>
    </value>
    <value>
        <val>5</val>
        <lab>Widowed</lab>
        <labor>Viudo(a)</labor>
        <freq>8803</freq>
    </value>
    <value>
        <val>6</val>
        <lab>Single</lab>
        <labor>Soltero(a)</labor>
        <freq>113123</freq>
    </value>
    <value>
        <val>7</val>
        <lab>NIU</lab>
        <labor></labor>
        <freq>96781</freq>
    </value>
</variable>
```

Figure 2. Part of Costa Rica 2000 Data Dictionary (untagged view)

Rec	Var	Col	Wid	Frm	Value	Va	ValueLabel	Va	ValueLabelOrig	Freq	Svar
					2		No		No Sabe leer y escribir	33,425	
					3		NIU			37,858	
P	marst	111	1				Marital Status		P13-Estado Conyugal		CR00A420
					1		Consensual union		Unido(a)	38,937	
					2		Married		Casado(a)	108,617	
					3		Separated		Separado(a)	8,975	
					4		Divorced		Divorciado(a)	6,264	
					5		Widowed		Viudo(a)	8,803	
					6		Single		Soltero(a)	113,123	
					7		NIU			96,781	
P	econact	112	2				Economic Activity		P14-Condición de Actividad		CR00A421
					1		Employed		Trabajó	126,085	
					2		Employed not paid		Trabajó sin pago	1,145	
					3		Didn't work but was employed		No trabajó,tenía	2,668	
					4		Looked for work having worked before		Buscó trab.había trabajado	5,519	
					5		Looked for work for the first time		Buscó trabajo 1ra. vez	841	
					6		Pensioner/Rentier		Pensionado/rentista	13,062	
					7		Studying and didn't work		Estudia no trabajó	44,863	
					8		Household duties		Trabajos del hogar	78,080	
					9		Other		Otro	12,456	
					10		NIU			96,781	
P	sector	114	1				Institutional Sector		P15-Sector Institucional		CR00A422
					1		Federal government		Gobierno Central	9,375	
					2		Autonomous Service Institution		Institución Autónoma Servicio	5,120	
					3		Autonomous Financial Institution		Institución Autónoma Financieras	1,602	
					4		Autonomous Not Financial Institution		Institución Autónoma No Financieras	2,141	
					5		Public company		Empresas Públicas S.A.	472	
					6		Municipalities		Municipalidades	918	
					7		Private sector		Sector Privado	110,032	
					8		International organizations		Organismos Internacional	238	
					9		NIU			251,602	
P	ind	115	2				Industry		P16a-Rama de Actividad a 2 digitos		CR00A423
					1		Agriculture, cattle ranching, hunting, and rel		AGRICULTURA, GANADERIA, CAZA Y AC	24,126	
					2		Forestry, wood extraction, and related servic		SILVICULTURA, EXTRACCION DE MADE	402	
					5		Fishery, fish raising and fish farming, relate		PESCA, EXPLOTACION DE CRIADEROS	678	
					10		Coal and lignite mining, peat extraction		EXTRACCION DE CARBON Y LIGNITO; E	0	
					11		Oil and natural gas, related service industrie		EXTRACCION DE PETROLEO CRUDO Y	3	
					12		Uranium and thorium minerals extraction		EXTRACCION DE MINERALES DE URAN	0	
					13		Metal mineral extraction		EXTRACCION DE MINERALES METALIFE	29	

format; other macros validate the metadata before it is exported to XML format. As we continue to develop metadata over the next several years, we will continuously expand the capabilities of the VBA macro library, to minimize the need for manual tagging, improve metadata quality, and increase production speed.

The scale of the metadata required for IPUMS-International is large. For example, for just the first 28 datasets, the metadata describing enumeration instructions for the labor force participation variable is the equivalent of over 100 single-spaced pages. Accordingly, the development of metadata is one of the primary tasks of the IPUMS-International work plan, as the following section makes clear.

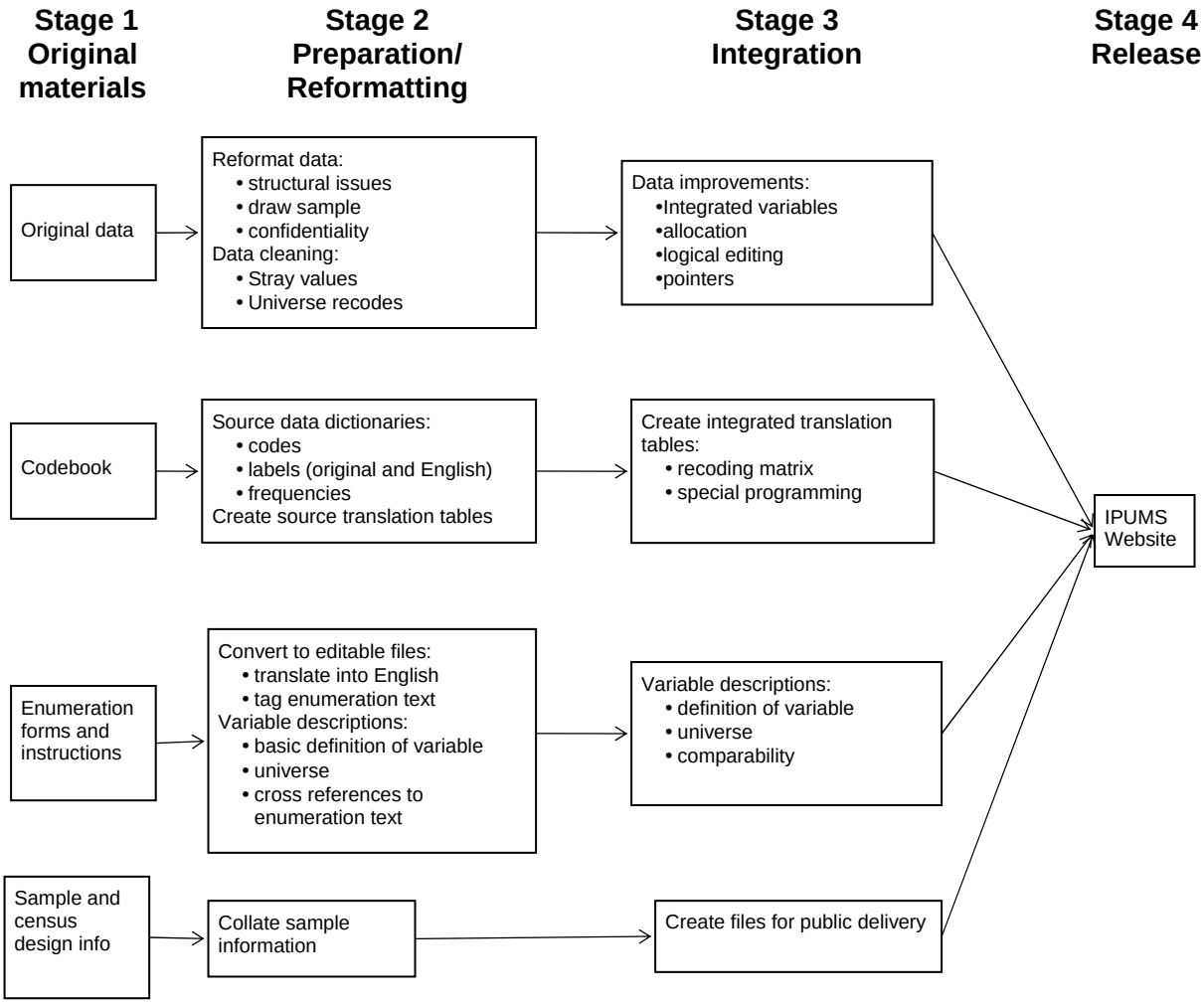
Work Process

In tandem with reconfiguring much of the underlying programming, we have completely redesigned our work process during the past year. This redesign was essential: we are ramping up the pace of production at least 500%, and the hand-crafted approach we took for the first release is impractical. The sections that follow summarize the major tasks associated with each component of the redesigned work process. Figure 3 provides a

generalized overview of the process, from the input materials provided by national statistical offices to the dissemination of the database on the web.

The data go through three major stages of processing. First, we assemble the raw data and documentation. Second, in the preparation and reformatting stage we reformat the data into column-format hierarchical ASCII files, draw a sample if necessary, and impose basic confidentiality edits as needed. In this stage, we also clean the data, eliminating stray values and creating separate categories for values not in the universe. Third, we integrate the data, which includes adding variables that are compatible across countries and census years, editing and allocating missing and inconsistent values, and constructing new variables to simplify analysis. Our data processing software is driven by metadata, so developing metadata is a central component of each phase of work. The sections that follow describe each stage of processing in turn.

Figure 3. IPUMS Process Schematic



1. Collect original materials

The first step in processing is to acquire from member countries the source materials (e.g., enumeration forms, instructions to enumerators, data dictionaries) needed to carry out all subsequent processing. Although not technically demanding, acquisition can be a significant bookkeeping challenge and may require considerable correspondence. Materials may come in numerous accessions with opaque filenames and often unclear content that may not be in English. Most material is sent to us via CD, but sometimes the transaction is entirely electronic or via paper documents (which must be scanned if relevant). All the files must be archived, systematically renamed to reflect their source and content, and organized in a workable directory structure on our network. We must also confirm that data dictionaries actually correspond to the data file(s)--which is never a given when decades-old material is pulled together by our overseas partners. Data files are sometimes unreadable or not drawn to the desired specifications. Information on complex sample designs is often particularly difficult to acquire. Depending on the responsiveness of a particular international partner, it can take weeks to months of sporadic correspondence to get what we need.

2. Preparation and reformatting

The second stage of processing is the most labor intensive. We must convert the data and metadata—received in various formats and languages—into systematic inputs for all subsequent IPUMS processing. At this stage our research staff uncovers and deals with many data and documentation errors and omissions.

Language translation. Most countries send their documentation to us in languages other than English. Maintaining a staff with all the requisite language skills is impossible, so these documents must be translated into English before processing begins. In the past, translation was sometimes done in an *ad hoc* manner, as required at particular stages in the work. Now, the demands of our new work process require timely and thorough translations. 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. Sometimes we hire translators from language departments at the University of Minnesota; in other cases, we identify foreign nationals to perform this work offsite. Our preference is for native English speakers, who generally translate into better English, but we have learned to be flexible. Apart from the expense, language translation is an administrative burden, because it is often carried out remotely, involves contracts, and can require hard-to-predict lead time as we try to schedule future data releases.

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 format easily readable by software. Most datasets we receive do not have unified dictionaries. Instead, several documents provide information on complex variables like occupation, industry, and geography. We pull this information together in the source data dictionaries discussed in the previous section. These metadata have features not accommodated in a typical codebook, including a record of all original-language labels side-by-side with their English-language versions for samples where translation was necessary.

Many of our tools used in data processing rely on these metadata. For example, we have a suite of utilities to check the validity of the XML-tagged data dictionaries. These tools help ensure that our metadata for a sample is sound from the outset of the process. Another set of utilities computes frequency distributions for every variable in the input data files and inserts these distributions into the data dictionaries and, later in the process, into the translation tables used for integration.. A third utility uses the data dictionaries to produce SPSS syntax files for the entire dataset or for selected variables and record types. Since we edit the variable labels at various stages of processing, statistical package syntax files can easily become outdated. This utility allows us to generate updated syntax files on demand.

A major innovation in the past year of development has been the formal incorporation of every original source variable from each dataset as its own unique IPUMS variable. In the past, source variables were used to create IPUMS variables, but they were not treated as IPUMS variables themselves. They did not have a complete set of metadata, and thus they were not easily accessible to IPUMS research staff through the data access software. Now, in IPUMS-International, source variables are treated like any other IPUMS variable: they are stand-alone, coherently-documented units, and they are also the building blocks for the later integration stage of processing. Moreover, in the new metadata system, the source variables provide the link for connecting the final integrated material back to the original source metadata (such as the census questionnaires and instructions).

We assign each original source variable a unique name within the IPUMS system. We have standardized these names; for example, CR84A023 is variable 23 from the 1984 Costa Rica sample. (Household variables start at 0, and person variables start at 400, so we can readily tell the record type from the name.) The "A" identifies it as sample A, to distinguish between more than one dataset (e.g., with different geographic identifiers) from a given census. The input or source variable names are generated by computer and do not change over the course of the project. If we subsequently discover the need to add additional variables for some purpose, we assign the next highest variable number, regardless of its position within the record structure.

The next stage of metadata processing involves associating specific enumeration text for each census with the individual variables in the source data. This work can commence when we have English language enumeration text and a data dictionary assigning each input variable a unique, formulaic name (e.g., CR84A023, above). The first step in associating metadata text with a source variable is to insert XML formatting tags into the enumeration forms and instructions so they will render properly on the web. We use a Visual Basic tool for this tagging work, so researchers need not type XML. The research staff identify every block of text that corresponds to a source variable and tag these blocks accordingly. That text is thereafter permanently associated with the source variable. The tagging lets us compile, on demand, all enumeration text for any single source variable or for all of the source variables that underlie any integrated variable.

Once we have tagged the enumeration materials, we know the text of the census form and the instructions to respondents and to enumerators (if any) associated with that specific source variables. To regularize the variables as proper input and to present them publicly,

however, we need a formal variable description. We write the description based on the tagged enumeration text.

Finally, we document sample and census design information. Some countries compile this information for us; in other cases, we must scan through the available documentation, ask our questions of our contacts, and possibly infer missing elements. Like all our metadata, this ancillary information is stored in marked-up structured documents that make machine-processing comparatively easy.

Data reformatting. Once the data dictionary has been defined and the source variables have been uniquely named, 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—need to be corrected. During reformatting, we add some basic IPUMS 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.

Before the redesign of our work process, data reformatting was done with a sample-by-sample approach. Over the last year, however, we have developed a flexible set of general-purpose reformatting tools to increase the efficiency and accuracy of the reformatting process. Some diagnostic tools are used before reformatting, to reveal structural problems that we must address during reformatting. Other tools are invoked within the reformatting program itself. The latter handle routine tasks common to most reformatting jobs—for example, reading all the input records that define a dwelling, computing the number of heads and spouses within households, checking for duplicate person records within a dwelling, and calculating serial numbers across a data set. With this toolkit, less of the analyst's effort is expended on the routine aspects of data reformatting, and more time is left to focus on especially challenging and truly sample-specific reformatting details. Finally, during the last year we also developed a flexible model (or template) for writing sample-specific reformatting, making work on each sample more efficient and robust.

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). For example, household records and person records are sometimes delivered in separate files, and occasionally there will be no household record corresponding to a set of person records. If there is no clear 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). Our software passes through the entire dataset, categorizing dwellings into strata defined by dwelling type and a set of predictor variables. On a second pass, when the software encounters a bad dwelling, it substitutes the most proximate potential donor within the same stratum. We substitute that donor dwelling while retaining

the geographic information from the original. The program prevents the repeated substitution of the same donor dwelling by maintaining a stack of dwellings available for donation within each stratum. Donor dwellings are identified with a flag. Prior to donation we carry out a dry run to ensure that each stratum has a viable ratio of good to bad dwellings and adjust the definitions of the strata if necessary.

In many cases we have full-count data or high-density samples that cannot be released as public use files. As part of our procedures for creating anonymized IPUMS-format files, we draw samples for public distribution. Working through a geographically-sorted file, we take a systematic sample of dwellings from a random starting point to yield what are typically 10 percent public-use samples. With respect to data errors, we use the same procedure described above for identifying bad dwellings and defining donation strata.

Very large units are sampled differently. Over time and across countries, the group quarters concept (referring to collective/institutional households and private households with a given number of unrelated persons) is applied inconsistently. Some collective dwellings have hundreds of records; some households that are clearly collective are not so identified. The standard errors on large collectives are large and can yield misleading statistics. Moreover, very large households, whether collective or private, pose potential confidentiality risks. For all these reasons, we impose a consistent maximum household size threshold across all samples in IPUMS-International.

We have adopted a threshold of 30 persons as the maximum for household-level sampling. In the datasets we have processed so far, households larger than this make up far less than one percent of dwellings. Very few private households over 30 persons in these samples have been genuine, with most such cases deriving their unusual size from structural data problems (e.g., the intermingling of two or more separate households assigned duplicate serial numbers). In several samples we have already imposed a 30-person limit, and we expect to do so in future. Under this practice, any household we encounter with more than 30 inhabitants is broken up and sampled individually, creating single-person group quarters units. In such cases,we indicate, with a data quality flag, that the cases are sampled from a larger unit and note the size of the original large dwelling. Collective households with 30 or fewer persons are taken as intact units. We will, however, consider modifying this sampling rule if we encounter populations in which a significant fraction of households exceed 30 persons.

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.

These confidentiality measures are imposed at the end of the reformatting and sampling stage. 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 also 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.

We retain a copy of the original unsuppressed and unswapped dataset, in case we need to return to it for some reason, (such as discovering, in light of a renegotiated country agreement, that we were overly aggressive in removing geographic detail.) This safeguard leaves open the door to later adding contextual information or doing other data manipulation that requires sensitive (e.g., low level geographic) information for processing. To protect subject confidentiality and fully honor our distribution agreement with each international partner, we then encrypt the unsuppressed version of the data and all earlier iterations. Only project senior staff have access to the encryption key.

Some confidentiality procedures are carried out after reformatting, during the data standardization phase. Generally these measures speak to lower-order confidentiality concerns. They involve recoding 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).

After reformatting, sampling, donation, and confidentiality edits, we create a new version of the data dictionary to reflect the final state of the input data. We update the data dictionary to incorporate changes in variables and the new frequency counts in the final sample. At the end of this stage, we have the processed input dataset that will be used for all subsequent work, and we archive the raw input data.

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.

We have developed standard procedures for performing this universe verification for source variables. Our research staff verify the universe in a two-way cross-tabulation, comparing the NIU (not in universe) category for each variable with a variable constructed to fit the stated universe. We then have them document the extent of Type I and Type II errors, respectively: persons not expected to be in the universe who have responses for the variable in question; and persons expected to be in the universe who are coded as NIU. We do not alter the data at this time, because we do not know which variable is incorrect--the one we are examining or the one that defines the universe. Errors uncovered during this investigative work are best resolved, we believe, through a process of missing data allocation and logical editing, which we expect to do in the future. For now, we are simply documenting where the problems exist.

We also perform some cleaning of the raw source variables as we document them. We put stray undocumented values into a unified "unknown" category, impose some basic rules about coding the NIU and "unknown" categories, resolve missing labels when possible, recode all alphabetic values into numeric codes, and generally rationalize and standardize coding. Finally, when the NIU category is combined with another, meaningful category (for example, when adults with zero income are combined with infants coded as 0 because they are NIU for the income question), we write an algorithm to disentangle the two categories. These algorithms form part of the variable-level metadata. We are careful not to lose any meaningful information as we recode and re-label the source variables. Our primary goal is to systematize and clean up the variables, to simplify subsequent processing as much as possible. We also see benefits in having more regularized source variables, since in most cases we plan to make them directly available to researchers.

3. 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 variables 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.

The basic metadata for data integration is the translation table (termed "data transformation matrix" in the IPUMS-International proposal to NSF). There is a separate translation table for each integrated variable; part of a translation table, stripped of tags, is shown in Figure 4. This translation table covers some of the categories of the IPUMS household relationship variable RELATE (relationship to head of household), and the selected view shows codes for Brazil in 1991 and 2000, France in 1968, Mexico in 1990, and the United States in 1960. The leftmost column of the translation table gives the standardized IPUMS code; the first digit of the IPUMS code provides a level of detail available in all datasets, and the additional/trailing digits provide detail available in only a subset of datasets. Beginning with the third column from the left and moving right, each of the other columns in the table represents a particular dataset, and each cell contains the code and label from the processed source data corresponding to the standardized IPUMS code.

For each integrated variable, researchers examine every sample to locate source variables corresponding to the concept in question. They insert each source variable name in a column of the translation table (not shown in Figure 4). Software then retrieves the source variable metadata (codes, labels and frequencies) and inserts it in the integrated translation table. Researchers then manipulate the individual input codes for each source variable to associate them with the appropriate IPUMS code. While doing this they have access to all of the relevant enumeration text. Like the codes and labels, this material is also compiled by computer, based on the source variables identified in the translation table. With all this material in front of them, researchers rearrange the codes for each sample to align with the corresponding IPUMS codes in the translation table.

Figure 4. Part of IPUMS translation table for RELATE

code	label	br1991a	br2000a	fr1968a	mx1990a	us1960a
rectype	RELATE	P	P	P	P	P
columns	Relationship to head of household	17=18	72=73	41	6=8	48=51
proj1		BR91A401	BR00A409	FR68A418	MX90A402	US60A423
2000	SPOUSE/PARTNER	2 = Cônjuge	2 = cônjuge, companheiro(a)		200 = Esposa(o) o compañera (o) {123961}	
2100	Spouse			2 = conjoint legitime du chef de menage {21.4597926342302}	201 = Esposa (o) {742}	201 = Spouse
2100	(Husband)				204 = Marido {1}	
2100	2nd/3rd wife					202 = 2nd/3rd wife (polygamous)
2200	Unmarried partner			3 = conjoint illegitime {0.532523400399875}	202 = Compañera (o) {30}	1114 = Unmarried partner
2210	Concubine				203 = Concubina (o) {9}	
3000	CHILD		3 = filho(a), enteado(a)	4 = enfant (du chef de menage ou de son conjoint) {38.1924753736065}	300 = Hijo (a) {424124}	301 = Child
3000	*Child*				301 = Hijo (a) {3070}	
3100	Biological child	3 = filho/a				
3100	(Son)					
3100	(Daughter)					
3200	Adopted child	4 = enteado/a			303 = Adoptado (a) {204}	302 = Adopted child
3200	*Adopted, n.s.*					*304 = Adopted, n.s.*
3300	Stepchild				302 = Hijastro (a) {1217}	303 = Stepchild
4000	OTHER RELATIVE					1001 = Other relative, n.s.
4100	Grandchild	8 = neto/a ou bisneto/a	5 = neto(a), bisneto (a)		608 = Nieto (a) {21925}	901 = Grandchild
4110	Great grandchild				609 = Bisnieto (a) {351}	1051 = Great grandchild
4120	Great-great grandchild				610 = Tataranieto (a) {2}	
4200	Parent/parent-in-law		4 = Pai, mae, sogro(a)	5 = ascendants {2.01376489381287}		
4210	Parent	5 = pai o mae			601 = Padre/Madre {4592}	501 = Parent
4210	(Father)					
4210	(Mother)					
4211	Stepparent				602 = Madrastra o padrastro {67}	502 = Stepparent
4220	Parent-in-law	6 = sogro/a			614 = Suegro (a) {2320}	601 = Parent-in-law
4300	Child-in-law	9 = genro ou nora			615 = Nuera o Yerno {6147}	401 = Child-in-law

Frequently, a variable will not mesh perfectly with the existing IPUMS coding structure. Sometimes new IPUMS codes need to be created or their labels altered. In other cases, more substantial changes are needed, and an integrated variable must be completely redesigned. When redesign of an existing integrated variable is impractical, we spawn a parallel variable that can accommodate the idiosyncrasies of the new sample.

Researchers sometimes encounter source variables that cannot be easily aligned with the categories of an existed IPUMS variable. In some cases, for example, information from more than one source variable is needed to identify categories in an IPUMS variable. In this case, researchers note any logical programming needed to supplement the basic recoding operation of the translation table in pseudo-code at the bottom of the table.[3]

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 samples and any changes in the codes. The comparability descriptions require particular care; IPUMS researchers 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 will be able to examine the associated enumeration text for any integrated variable. In the future, users will also be able to examine the constituent source variables that served as input to the integrated version.

Metadata from the translation tables drives the IPUMS Data Conversion Program (DCP). which reads the reformatted, confidentialized, sample data and writes IPUMS-coded data. The program is written in C++. During the past year, our programming staff has redesigned this software to be driven by XML metadata rather than fragile parsers. The program operates from the same metadata as the web dissemination software, so there is no possibility of the two getting out of sync with one another.

In addition to producing integrated data files, the DCP generates a companion SPSS syntax file to read the output. Once again, the syntax file is created on the spot from the same metadata that drive the system, so it always matches the data. Each time we run program, the previous version of the data is automatically moved to an archive directory. The DCP creates flag variables that allow easy comparison of the old and new versions of the sample: each variable gets an accompanying flag that indicates if any values changed between data runs. The flag variables allow researchers to quickly ascertain if the changes are as expected and to detect inadvertent errors as soon as they are introduced. This is a valuable new addition, because, as the datasets grow and become more numerous, the burden of quality-checking steadily increases.

[3] At present, programmers must incorporate this code into the IPUMS data conversion program, but in the future it will be directly interpreted from the pseudo-code in the translation table. This change will increase efficiency, reduce errors, and allow analysts to test the effect of supplemental programming more easily.

Another feature added to the DCP during the last year is the computation of frequency counts for every variable. These frequency counts serve two purposes. The first is diagnostic. In the past, whenever MPC researchers ran a dataset through the DCP, they would have to analyze the output data using a statistical package to obtain frequencies; now, such information is immediately available. The information is written by the DCP to an XML file, and a utility assembles the frequency information for all samples into a single report. This report allows researchers to compare frequency distributions across all samples for any variable, quickly revealing samples with unusual distributions or unexpected output values.

The second purpose of the frequency metadata generated by the DCP is to drive our web data access system. Most variables have a codes page that display the variable's coding structure and indicates whether a specific code is available—and with what frequency—in a given sample. In the past, these codes pages were created by a separate computational pass over the data. Now, however, the codes pages are constructed from the XML frequency files created by the DCP. This not only increases efficiency by eliminating the need for a second time-consuming pass over the data but also ensures that the codes pages and the output data are always in sync.

4. Data release

Once we have converted the data to IPUMS format, dissemination is automatic. The data and documentation access software use the metadata developed in the previous steps, so virtually no additional work is needed. Over the past year, we have redesigned the most important parts of the IPUMS website to run using Java software driven by our XML metadata. These Java elements include the main variables page, the variable description and codes pages, and the data extraction system. If a user now clicks on our main variables page or a variable description, a Java program builds an HTML page on the fly from the metadata.

Because the software constructs pages dynamically, we can begin filtering the content that is put on the screen based on user-defined preferences. For example, users will have the option of specifying at the outset that they only wish to see variable availability and documentation from samples drawn from a particular region (e.g., from Europe and North America only) or dating from a particular period (e.g., 1990 and later). For the May release, we expect to demonstrate our first implementations of these capability on the variables page and in the variable-specific enumeration text. In future releases, we will add the additional data access features described in the grant proposal.

Staffing

The project employs 14 full or part-time staff, in addition to paying for smaller portions of time from Principal Investigators. There are 2.5 full-time research associates on staff. In Summer 2005 we hired five graduate research assistants, and between October 2005 and January 2006 we added three post-doctoral associates. We also have three full-time programmers. In addition, we hire language specialists as required to translate documents into English, and we typically have one or more undergraduates performing scanning of original documentation. We are in the process of hiring a full-time Project Coordinator

who will be responsible for handling much of the paperwork, budgetary oversight, and correspondence associated with the project. Now that we have redesigned our software and work procedures, we can increase staffing productively; we plan to add several additional researchers (graduate assistants, postdocs, or research fellows) for at least the next three years.

Work Schedule

In the first IPUMS-International project, datasets were processed in batches of five to ten samples at a time. For the expansion, we proposed working with all samples simultaneously. Our plan was to release subsets of variables for virtually all samples. Our logic was that this would improve the design of integrated variables, since all coding variations would be known from the outset.

That plan, however, proved to be unfeasible; the extensive preparation work required for each sample—translation, metadata development, cleaning, reformatting, and so on—would create a great bottleneck that would prevent us from releasing *any* data in a timely fashion. We therefore returned to the system of processing datasets in batches. We have, however, greatly increased the size of the batches; for each bi-annual data release, we are simultaneously processing 15 to 20 samples for a group of 5 or 6 countries.

The main disadvantage of this approach is that it necessitates the reconsideration of prior variable harmonization schemes. As the project has evolved, however, we have come to realize we will never have the full universe of world classifications at our disposal up front for design purposes. It is best to acknowledge the inevitability of variable redesign and to capitalize on an alternative work approach that yields considerable logistical advantages. In particular, the batch approach is consistent with systematized work processes that involve more people and get data out faster.

Our schedule is organized around biannual deadlines in May and November of each year. An overview of the schedule appears in Figure 5. The specific samples associated with particular data releases are subject to change, due to data or documentation difficulties or strategic considerations.

May 2006

Data release. 19 samples: Chile, Costa Rica, Ecuador, Venezuela, South Africa.[4]

[4] This set of countries has a large constituency. Requests for future data acquisitions indicate high demand for additional Latin American census microdata. Seventeen percent of requests mentioned Latin America, with data from Chile, Ecuador, Venezuela, and Argentina being the most desired. Sixteen percent of applicants requested access to more African data, with material from South Africa being the most sought-after.

Figure 5. Summary of IPUMS Processing Schedule

Deadline			Data release	Associated items	Batch 1	Batch 2	Batch 3	Batch 4	Batch 5	Batch 6
	Jan	2006			reformat, process	pick Nov 06 samples	pick May 07 samples			
	Feb	2006			process	translate, get docs	translate, get docs			
	Mar	2006			integrate	translate	translate	pick Nov 07 samples		
	Apr	2006			integrate	translate, reformat	translate			
NSF board mtg	May	2006	19 samples (Latin Amer, S Afr)	instructions/faq page	release	translate, reformat	translate, reformat			
Paris Euro mtg	Jun	2006		new registration system		process	translate, reformat	translate, get docs		
	Jul	2006		linked enumeration materials		process	reformat	translate		
	Aug	2006				process	reformat	translate		
	Sep	2006				integrate	reformat	translate		
	Oct	2006				integrate	process	translate		
	Nov	2006	~18 samples (Europe, Africa)	dynamic doc pages/user preferences		release	process	translate, reformat		
	Dec	2006		general and detailed variables			process	translate, reformat		
	Jan	2007		IDs for extract replication			process	reformat		
LA continuation	Feb	2007		standardized variables available			integrate	reformat	pick Nov 08 samples	
	Mar	2007		pointers for latest samples (possible)			integrate	reformat		
	Apr	2007					integrate	process		
	May	2007	~18 samples (Latin Amer, Asia)	upgrade original 28 samples			release	process		
	Jun	2007		advanced extract features				process	translate, get docs	
	Jul	2007		expand harmonized variable list				process	translate	pick May 09 samples
	Aug	2007		java data conversion program				integrate	translate	
	Sep	2007						integrate	translate	
	Oct	2007						integrate	translate	
	Nov	2007	~18 samples (LA, Europe, Africa)	pointers for latest samples				release	translate, reformat	
	Dec	2007		socioeconomic variables					translate, reformat	translate, get docs
	Jan	2008		on-line tabulation					reformat	translate
Eur continuation	Feb	2008							reformat	translate
	Mar	2008							reformat	translate
	Apr	2008							process	translate
	May	2008	[skip a release here or elsewhere]	missing data allocation					process	translate, reformat
End LA grant	Jun	2008		document data transformations					process	translate, reformat
	Jul	2008		data/census quality assessments					process	reformat
	Aug	2008							integrate	reformat
	Sep	2008							integrate	reformat
	Oct	2008							integrate	process
	Nov	2008	~18 samples (Asia/Africa, Europe)	variance estimation discussion					release	process
	Dec	2008		mirror site						process
	Jan	2009								process
	Feb	2009								integrate
	Mar	2009								integrate
	Apr	2009								integrate
	May	2009	~18 samples (Asia/Africa)	pointers for remaining samples						release
	Jun	2009		missing data alloc for latest samples						
	Jul	2009								
End Euro grant	Aug	2009								
End NSF grant	Sep	2009								

Instructions page. We will develop an instruction/FAQ page that will serve as a resource for experienced users and a logical point of entry for new users. We have compiled an extensive list of topics for inclusion. Among many other things, the FAQ will suggest how best to use the extract system and note the strengths and limitations of the data series. We will reorganize the "project description" section of the left navigation-bar to highlight the FAQ, which we expect will undergo continuous revision and amplification over the course of the project.

New registration system. A new user registration system will bring changes to both the application form and the underlying database that holds user-level information. We have designed a new application page that contains stricter language to protect respondent confidentiality. The new form also requires additional information about applicants and more extensive research descriptions, enabling more rigorous vetting on our part. Registrations will expire after one year and be renewable.

The database behind the registration system will not only control access, but will also serve as a vehicle to record user preferences for navigating the website. For example, users who are only interested in Brazil will have the ability to filter relevant web pages to contain only material appropriate to Brazil. We will add this user-preference functionality to the system in a subsequent data release.

Linked enumeration materials. All enumeration forms and instructions are translated into English and formatted consistently for textual web display. As describe above, we mark up the enumeration materials using XML tags that associate specific sections of the text with all relevant variables in the corresponding microdata sample. Users viewing the documentation for an integrated variable will be able to view the relevant enumeration text pertaining to the underlying census question(s) for any or all samples. They will also be able to view any particular piece of enumeration text within the context of its parent document. Finally, images of the original documents will also be available, allowing users to view the enumeration forms and instructions in the original language.

Dynamic main variables page. The main page listing the availability of variables across all the samples in the data series is the logical starting point for most IPUMS users. It suggests what kinds of analyses are possible and whether they can be performed cross-nationally. It is therefore necessary to be able to see all of the variable information for exploratory purposes, but it is also important to have the capacity to winnow the page, which will become very large with this data release. We do not expect to have a global user-preference system in place at this time, but we are designing a page that allows users to customize the view of this particular page, limiting the table to specific samples of interest.

November 2006

Data release. 20 samples: Belarus, Greece, Romania, Spain, United Kingdom, Phillipines, Cambodia, Uganda, Canada.[5]

[5] Over 100 of the 761 requests for future data acquisitions specifically mentioned data from the UK and/or Canada. Demand for European census data generally is high (almost one-third of requests),

Pointers for latest samples. Each census includes data on households and the relationships of individuals within households. To facilitate analysis, we create individual-level variables describing interrelationships among family members. The most important of these are three pointer variables that give the person number within the household of each individual's own mother, father, and spouse. These variables help researchers create measures of kin characteristics, fertility, marriage patterns, and family composition that are tailored to their specific research questions and analytic strategies. We also provide other fully compatible variables describing family and household characteristics at the individual and household level. These include family membership, family size, number of own children, number of own children under five years old, and age of eldest and youngest own children.

The development of family interrelationship pointers is separable from the release of the basic data for each sample. The pointers are constructed variables linking spouses to one another and children to parents. The work is complex, and it is not efficient to make this task part of every data release. Instead, just prior to the November, 2006 data release, we intend to add the pointer variables to all samples from previous releases that lack them (approximately 28 datasets). We will also modify the existing pointers to conform to any evolution in the linking algorithms. During the first IPUMS-International grant period, we were conservative and limited parent-child links to persons under age 19. During this upcoming period, we will explore the feasibility of removing that restriction.

As part of this pointer development, we will also attempt to systematically identify the full range of logical functions necessary to construct the links, with the goal of partially automating the process in the future. In all previous data releases, going back to the original IPUMS-USA, the pointers have followed a core set of age, relationship, and positional rules, but we have customized pointer variable creation to accommodate differences among samples. Up to this point, the logical rules have been worked out by researchers; the researchers have transmitted the information as text and pseudo-code to programmers; and the programmers have implemented the work via the data conversion program. For basic variable recoding, our new model implies a shift to scripts, written by researchers, that are read directly by the data conversion program. If we can isolate the basic set of necessary functions for pointer variables, we will try to do the same thing for these constructed variables. Removing the programmers from the work process would be highly advantageous from an efficiency standpoint. Nonetheless, this might prove unfeasible because the range of required functionality of even a basic set of tools might be too great. In any case, the pointer variables constructed just prior to the November 2006 data release will be developed much as in past releases, albeit with an eye toward overhauling the method next time.

Because our timetable will still be compressed through the November release, it is uncertain whether we will fully complete the pointer variable component on schedule. Pointer development is highly skilled work, but other critical processes may consume all

and there is also expressed demand for both additional Asian data (one quarter of requests) and additional African samples (16 percent of requests).

available time from the project's Research Associates. The ability to complete the pointers for this release may depend on the suitability of one or more of the other research staff for such work.

User preferences—dynamic variable pages. Virtually all of our web pages will be generated from underlying documents we mark up with XML tags. In the case of variable lists, variable descriptions and a number of other pages, the tags note the country or sample to which a discussion or statistic pertain. Thus, with the proper input we can customize many web pages on the fly to correspond to user preferences. In the previous release, we customized the main variables page, which was proving unwieldy. By this data release, more pages need to be customizable to avoid overwhelming users: specifically the harmonized variable descriptions and codes pages. The system will limit the information presented on these pages to the choices expressed by the user. Users will be able to select countries and samples at any point in their session and can store those preferences for subsequent sessions. As we develop this functionality we will consider the options for substituting defined preferences with clickable points on the web page (for example, collapsing and expanding trees). This would require adding a technology such as Javascript to the dissemination system.

General and detailed variables. Most IPUMS harmonized variables have a multi-digit coding scheme designed to identify which levels of information are largely comparable across samples and which samples provide greater detail within a larger category. Home ownership, for instance, has only two fully-comparable categories at the first digit—owned and not owned—but has second and third digits that provide more detail for specific samples. At present, the hierarchal nature of the coding schemes is not explicitly recognized by the IPUMS-International software; users cannot simply choose the simplified version of a variable and eschew the extra detail. We suspect that the complex coding schemes often prove an impediment for users, most of whom either don't require the maximum detail or need such detail for only a few variables.

IPUMS-USA now offer users a choice between general and detailed versions of many variables. IPUMS-International's current metadata design already anticipates the general variables feature. However, a number of web programming features must be added to IPUMS-International to make general and detailed variables appear uniquely in the main variable listing and the data extraction system, and to handle the statistical package syntax. We must also add a new field to the variable descriptions to identify which sections of the discussion apply to the general versus the detailed version of the variable.

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.

Processed source variables. In the process of adding new samples to the IPUMS-International database, all of the original input or source variables are individually documented. As described earlier, to facilitate subsequent international harmonization and to ensure that we fully understand each variable, we process the source variables to some

extent. Specifically, as needed, we recode the source variables to clean up stray values; we split off NIUs when they are conflated with other categories; and we write brief (single-sentence) descriptions.

With so much processing already carried out to make the source variables usable for our internal processing, the next logical step is to make them publicly available through our data access system. The work required to do this is largely an issue of web design—albeit a substantial one—regarding how to present the standardized source variables on the variable list and in the data extraction system. A fundamental question is whether we allow users to access source variables directly by sample or only through the prism of specific integrated variables.

The availability of the processed source variables presents several opportunities. First, they would allow us to modify our general strategy regarding integration, focusing the intellectual work of integration on important common variables. Unique and incompatibly-coded variables would be accessible to the public, but only through these processed versions of the original variables. Second, the existence of the processed source variables also might give us the latitude to "lose" some information in an integrated variable, when the inclusion of some details overcomplicates the coding structure. Such losses in an integrated variable would be tolerable, because the full detail would still be available to researchers via the underlying standardized variables. Finally, we could use the processed source variables as a means of releasing a sample's data prior to harmonizing it, thereby giving researchers faster access to the sample.

May 2007

Data release. ~20 samples: Bolivia, Panama, Puerto Rico, Peru, Israel, Malaysia.

Java data conversion program. The current data conversion program that converts input samples into IPUMS data is written in C++ and has a number of features we would like to change. For numerous technical and practical reasons, the MPC IT core has settled on Java as the language of choice for this sort of application. A member of the IT programming staff will therefore rewrite the IPUMS data conversion program in Java, while simultaneously making a number of improvements. This significant overhaul will absorb much design time but is necessary for the long-term maintenance of the system.

The principal new innovation, from the research perspective, will be the addition of a scripting feature, as mentioned above. Henceforth, for most variable recodes that require programming, the researchers will write the code in a scripting language that will be read directly by the data conversion program. The programmers will no longer serve as middlemen in the process. This should enhance researcher control over the final product, reduce communication-related errors, and facilitate data fixes.

Expand integrated variable list. We will review the integrated variables from previously-released samples. Depending on the role we settle on for the processed source variables, we may expand the number of integrated variables. This task may be spread out over more than one data release, as needed.

Upgrade the original 28 samples. The original 28 samples for 8 countries, developed under the first IPUMS-International grant, were subject to a different design regime. We have retrofitted certain aspects of our current methods back onto those samples, like the organization of the metadata. Still, a number of data improvements need to be made to ensure consistency and uniform practices across the older and newer samples. We will probably choose to undo some of the logical edits we implemented to yield "clean" universes in these samples. Most important, we need to revisit data reformatting, treatment of structural data problems, and handling of large dwellings. The underlying data will change for some samples.

Advanced extract features. We construct only a limited number of variables during IPUMS data sample creation, but the hierarchical structure of the data offers tremendous opportunities in this regard. We will make it possible for researchers to construct a variety of descriptive variables tailor-made for their individual research questions, 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.

November 2007

Data release. ~18 samples.

Pointers for latest samples. We will add constructed family interrelationship pointers to the samples processed since the November 2006 data release. Most significantly, if feasible, we will implement a scripting language for the pointers at this time. This will introduce greater consistency and researcher control over the pointers and make their development less burdensome.

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.

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 a fair amount of work involved developing a web interface suited to our microdata. As discussed in a later section on Education and Outreach, the on-line tabulation utility may attact less sophisticated data users and prompt

new outreach and training efforts (e.g., summer workshops for teachers of high school and undergraduate students).

May 2008

No data release. We expect to skip one data release to allow for infrastructural and methodological improvements. At this time, we have scheduled the break for this release date, but it may become beneficial to skip one of the nearby data releases instead.

Missing data allocation. By "missing data allocation" we mean both probabilistic imputation and logical editing of records. We will not carry out allocation on all variables. Instead, we will focus on the variables which are most used by researchers and most likely to generate logical inconsistencies. A significant portion of our effort will go into the editing of the age and relationship variables. This editing is complex because the responses in these variables must be consistent with each other, with other information on a person's record, and with other persons in the household. The potential for conflicting information varies by sample. By editing the age variable, we will be able to clean up universes of many other variables, yielding a more user-friendly product for researchers.

We are likely to revise the allocation and editing procedures for some or all samples as we refine our methods late in the project. As we apply the first-generation allocation software to the international data, we will evaluate the possibilities for introducing more flexible scripting functions, to enable greater researcher control over data editing. If this proves practical, we will be able to do more comprehensive data editing across the samples.

When we allocate or edit the data, we will indicate the altered records with appropriate data quality flags. The flags will be made available through the data extraction system on a separate selection screen.

Document data transformations. Most data transformations are documented in the integration tables, including how input data values and labels correspond to IPUMS values and labels. We will determine the most practical format for delivering this information to users on the web. It may be possible to link users to the source variables through this device to further enable researchers to explore and even deconstruct our coding decisions. 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 as we convert to the Java data conversion program. 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.

Data quality assessments. Researchers will attempt to derive indications of census data quality for each sample to be included in our web documentation. We will compare selected results from the sample data to published statistics; search the academic literature for under-enumeration estimates for the various censuses; and carry out statistical analysis of the data, such as evaluating the degree of age heaping. To the extent feasible, we will summarize results of post-enumeration surveys carried out by national statistical offices.

We will be unable to carry out these steps for many samples because of the difficulty obtaining the necessary documentation. In particular, post-enumeration studies are not widely accessible, and the costs of translation into English may be prohibitive.

November 2008

Data release. ~18 samples.

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

Mirror site. The IPUMS web site will be mirrored at ICPSR. We are currently working on mirroring the IPUMS-USA system. Both the USA and International systems are driven by the same metadata design, so it should be easy to set up mirroring for IPUMS-International. The main reason for mirroring the site is to ensure preservation of the complete data, metadata, and software system. The mirror site will also provide for rapid restoration of service in the event of a problem with the Minneapolis server. Until such a problem arises, however, we intend to disseminate data exclusively through the Minneapolis server.

May 2009

Data release. ~18 samples.

Pointers for remaining variables. We will develop family interrelationship pointers and associated constructed variables for all remaining samples.

Missing data allocation for latest samples. We will perform missing data allocation for all remaining samples. This work will include both performing missing data allocation for the same set of variables as earlier datasets and performing missing data allocation for any additional variables deemed necessary and feasible.

Security and Preservation

Only staff who work on the project and senior MPC supervisors have read access to the IPUMS-International work area on the MPC network. All persons who do have access must sign agreements not to make copies of the data and not to attempt to identify individuals in the data. The alpha website where we develop the data and test various web features is password- protected. As noted above, non-anonymized data are encrypted, and only senior project staff have the encryption key. Only samples that have passed through our confidentiality measures are accessible to other IPUMS-International staff.

We are also improving the data security agreement for external users; a new registration form commits researchers to more stringent requirements. The user application form also asks for more information that allows us to verify the identity of the applicant.

The biggest security risk does not come from unauthorized use but from authorized users redistributing the data to colleagues, posting it on a website, or storing it in an unsecured

area, none of which are allowable. We can, however, readily identify unauthorized distribution of IPUMS datasets and bring sanctions to bear. All IPUMS-International data have been subjected to distinctive data processing, and the data available from our website are easily distinguishable from the originals because of the standardization and reformatting we perform on all datasets.

As our collection of data grows, and more and more of the files are uniquely held at Minnesota, the issue of preservation has become a priority. All data and electronic documents we receive are stored on the MPC network. Any original physical media are stored on-site. Original data and electronic documents are also copied to another set of physical media (CDs and DVDs) and stored off-site in Minneapolis.

The entire IPUMS-International network area, including all metadata, programs, and data, are incrementally backed up daily on the Center's RAID array. Complete tape back-ups are made weekly and periodically stored off-site. Every three months tapes are sent to a long-term secure storage facility (Iron Mountain). The website is backed up as well, and will soon be cold-mirrored in its functional entirety at the Inter-university Consortium for Political and Social Research (ICPSR).

Data Acquisition

We have pursued energetic data acquisition policies from the beginning of the project in 1999, and our efforts have paid off beyond our most optimistic initial hopes. By the end of 2005, we had archived the largest collection of census microdata in the world, with datasets for 140 censuses representing 50 countries. The details of negotiations and acquisitions by region, country, and census appear in Appendices A and B.

We expect 2006 to be our most successful year yet, thanks to major breakthroughs with statistical agencies in Africa. Such breakthroughs have come about as a result of recent negotiations with official statisticians at conferences in Cape Town (September 2005 and January, 2006), Kampala (November, 2005), and Addis Ababa (February, 2006).

Our efforts in Africa were initially delayed as we sought to negotiate a mutually beneficial arrangement with the African Census Analysis Project (ACAP) at the University of Pennsylvania. Finally, in 2005, with the encouragment of the Statistican General of South Africa, Pali Lehohla, we decided to initiate direct negotiations with African statistical agencies. We thereby broke a five-year long hiatus but left open the possibility of future collaboration with ACAP.

Considerable diplomacy is required to persuade official statistical authorities not only to entrust their data to a relatively little-known entity but also to cede to us the responsibility of determining who is to be given access to the microdata. Crucial to our successes are the blessings of senior international statisticians and demographers, some recently retired, all working without compensation. With their help, we have extended invitations to participate in IPUMS-International to most of the national statistical agencies of the world, including nations that have poor relations with the United States, such as Cuba, Iran, DPR Korea, and Uzbekistan.

For the moment, we are not actively pursuing data from a small number of exceedingly reclusive nations (e.g., Bhutan, Brunei, Djibouti, Haiti, Myanmar, Somalia, United Arab Emirates) and a number of islands with very small populations. In some cases, persistent efforts over more than five years have yet to yield tangible returns (e.g., Algeria, Australia, Finland, India, Republic of Korea, Japan, New Zealand, Norway, Thailand, and Ukraine). In other cases, breakthroughs either appear imminent (Argentina, Bangladesh, Ethiopia, Ghana, Indonesia, and Nigeria) or have occurred very recently (e.g., Indonesia, which has agreed to share data from three decadal censuses and two intercensal microcensuses).

Our strategy of requiring a signed memorandum of understanding between the University and each National Statistical Authority initially slows the pace of data acquisition. A formally endorsed agreement, however, provides an essential legal foundation for microdata acquisition and dissemination. The tenure of official statisticians is often brief; thus far, our agreement has withstood the test of changing leadership of our national statistical agency partners.

International meetings of official statisticians offer the best opportunities for negotiating data sharing agreements; persuading agencies to locate and preserve their at-risk older census microdata; and collecting data and documentation from countries who have formally signed agreements with MPC. The meetings also allow us to and strengthen ties with agencies once dissemination of the integrated microdata is underway.

At such meetings, we staff a booth, present invited papers, and discuss questions and concerns of official statisticians, researchers, and others. The fact that we continue to be invited back to these exclusive events is an indication of our success in establishing trust and credibility. Over the past twelve months, we were invited to and attended major regional official statistician meetings for Asia (Seattle, March), Latin America (Santiago, June), Europe (Geneva, November), and Africa (Addis Ababa, February), as well as the International Statistical Institute meeting in Sydney (April).

During the coming four years, we will continue to negotiate with national statistical agencies to preserve and disseminate census microdata. At our current pace, we anticipate that we will be able to preserve a large proportion—probably most—of the world's surviving microdata before the project ends. Our success may mean that our budgeted data acquisition funding will prove insufficient, and this may require either raising additional money or shifting funds from data processing to data acquisition.

We will also seek new agreements with countries that are already partners. Many agencies (including those of Sudan, Colombia, and dozens of other countries), are entrusting entire microdatasets to the MPC rather than samples. With appropriate safeguards for protecting confidentiality, these complete count datasets offer the potential for extraordinary new research opportunities. To offer access to these data, even within the most secure data enclave, we will need to execute new agreements. We will also need to negotiate agreements to obtain the next round of censuses. The 2010 round of censuses will provide a significant test of the sustainability of extending the IPUMS-International collaboration into coming decades.

Education and outreach

To coordinate activities with our partners around the world, stimulate new partnerships, and promote use of IPUMS-International data, we have organized a series of regional symposia and workshops. The first of these was held in January 2004 and involved 35 statisticians and demographers from the census departments of 15 Central and South American countries and regional organizations. Organized in conjunction with the Centro de Investigaciones sobre Dinámica Social (CIDS) of the Universidad Externado de Colombia, the symposium featured in-depth discussions of issues relating to Latin American census data. Over two days of meetings, participants learned from officials at DANE, the UN Economic Commission for Latin America and the Caribbean (ECLAC), and the Brazilian Institute of Geography and Statistics (IBGE) about their first-hand experience in working toward data harmonization. The participants also discussed issues relating to confidentiality and data commercialization, contributed to presentations on specific data concerns and issues of each country, and discussed the MPC's plans and procedures for the project.

In March 2005 we held a one-day workshop in Seattle, Washington in conjunction with the 22nd Population Census Conference of the Association of National Census and Statistical Directors of America, Asia and the Pacific (ANCSDAAP). IPUMS co-investigators Robert McCaa and Dennis Ahlburg joined representatives from Asian and Pacific statistical agencies—including Cambodia, Fiji, India, Indonesia, Iraq, Japan, Malaysia, Mongolia, and the Philippines—along with a delegation from the Institute of Statistics of Chile and numerous participants from the Center for Studies in Demography and Ecology at the University of Washington. Workshop topics included preservation, anonymization, and dissemination, as well as temporal and geographic harmonization. Professor Ahlburg presented his work on Pacific Islanders in the United States as an example of applied research using IPUMS data; René Saa Vidal from INE Chile presented the possibility of a Geo-Demographic Information System based on harmonized census microdata. The Minnesota Population Center paid the expenses for representatives of several countries to attend the meeting on the condition that they deliver data to IPUMS-International by the time the meeting began. That incentive worked well, and resulted in significant new data acquisitions.

In July 2005, we held a three-day workshop in Barcelona, Spain organized with the Centre d'Estudis Demogràfics (CED) of the Univeristat Autònoma de Barcelona. Representatives from 14 European statistical offices met with IPUMS-International investigators to discuss plans for census data harmonization. Participants described the state of the census samples available from their country, highlighting potential issues for successful data harmonization. Several non-European countries also attended the meetings and contributed their data to the wider IPUMS project. The meeting demonstrated the keen interest within the European statistical community in the IPUMS project and helped build momentum for our collaborations. The meeting strengthened our European partnerships and allowed us to clarify our objectives and methods. Moreover, we learned a great deal about idiosyncrasies of European datasets. We also gained a new appreciation of the desire among the European representatives for ongoing engagement with the project, and the need to develop regionally optimized components of the IPUMS system.

To help us coordinate activities with our European partners and to develop new variables optimized for Europe, we have acquired a strong partner in the CED. Albert Esteve, a former post-doctoral associate on the IPUMS project and now a CED researcher, has received three European Union grants to further the work of the project in Europe. With the strong support of Anna Cabré, the center director, the CED will hold workshops, develop European-specific variables, collect supplementary data and documents of specialized regional interest, and serve as a facilitator in our relations with the various statistical offices. The CED will also house a mirror site of the IPUMS system with a customized front-end emphasizing the European focus.

A follow-up meeting with European statistical agency representatives, funded largely with European research monies, is scheduled for June 2006, in Paris. We have invited representatives from other regions, particularly Africa, to attend, conditional on their producing signed agreements and turning over the necessary data and documentation by the time of the conference. We believe that via this incentive mechanism, as many as 10-12 additional countries will sign on by early summer.

The strategy of regional meetings workshops, and symposia focused on IPUMS-International has proven highly successful, and we plan to continue to organize similar meetings. We are exploring the potential for a workshop held in conjunction with the April 2007 ANCSDAAP meeting in Christchurch, New Zealand. In addition, Statistics South Africa has invited us to co-organize a meeting in conjunction with the August 2009 meeting of the International Statistical Institute in Durban, South Africa.

Whether or not we organize formal workshops, we plan to attend every international meeting of statistical agencies that we can. As described in the last section, these meetings play a crucial role in our data acquisition strategy. They also, however, play an important role in education and outreach, by publicizing the availability of data. These meetings can foster capacity building by sparking discussions across national and institutional boundaries between the users and producers of national census data. IPUMS-International will have a strong presence at many upcoming international gatherings, including the Third Pan-African Census Symposium in Rwanda (January, 2007) and the meeting of the Asian and Pacific Population Group organized by Statistics New Zealand in Christchurch (April, 2007).

Academic meetings are also important, since they are key venues to inform potential users of the availability of IPUMS-International microdata through exhibits, papers, and mini-workshops. To publicize its data products, MPC routinely maintains an exhibit booth at several academic and professional conferences every year. These conferences include the meetings of the Population Association of America, the American Sociological Association, and the American Economics Association, the International Union for the Scientific Study of Population (IUSSP), the International Statistical Institute (ISI), and the International Institute of Sociology World Congress. Regional meetings, such as the Latin American Population Association (ALAP), European Association of Population Studies (EAPS), African Population Studies (APS), will become increasingly important as our regional offerings expand. We will have a strong presence at the upcoming EAPS meeting in Liverpool (June) and ALAP meeting in Guadalajara (September).

At most meetings, MPC rents exhibit space shared by IPUMS-International and other data products of the Minnesota Population Center. MPC staff hand out brochures, answer questions, and demonstrate the features of IPUMS-International data and website on a laptop computer. Because a single exhibit booth handles publicity for a host of data projects (e.g., IPUMS-USA, the Integrated Health Interview Survey, the North American Population Project, the National Historical Geographic Information System, as well as IPUMS-I), this approach is cost effective. Through individual conference presentations and public exhibits, MPC staff also periodically publicize our data projects, including IPUMS-International, at other U.S. professional meetings (e.g., the Joint Statistical Meetings, the American Historical Association, the American Library Association, the American Public Health Association, the Association of American Geographers, and the International Association for Social Science Information Service and Technology).

Increasingly, we conduct workshops to introduce users to our datasets. The workshops attract 25 to 35 participants, led by Dr. Trent Alexander (Director of user support services for MPC data projects) and Dr. Matt Sobek (IPUMS-International Project Coordinator). We have held half-day workshops in conjunction with the meetings of the American Sociological Association that provide training with IPUMS-USA and IPUMS-International data. We have offered two versions of this training, one of which focuses on research, and the other on undergraduate teaching. We conducted workshops at the ASA meetings in 2003, 2004, and 2005, and another such workshop is scheduled for the 2006 ASA meetings. We have also conducted day-long workshops on using MPC data sources, including IPUMS-International, in conjunction with the ICPSR Summer Program in Quantitative Methods, held in Ann Arbor, Michigan.

As discussed at the first meeting with the IPUMS-International Advisory Board, we are expanding our training efforts this year. Specifically, we are offering a 3-day IPUMS Summer workshop, scheduled for July 19-21st at the Minnesota Population Center. This workshop will teach social scientists how to use data from IPUMS-USA, IPUMS-International, IPUMS-CPS, and the North American Population Project. Together, these MPC-produced datasets cover 150 years of U.S. and international census data. Each day's training session will consist of a combination of presentations and laboratory work. The course will cover sample designs, database creation, data extraction systems, and complex issues involved in using the data (such as weights, geographic variables, and SES measurement). The workshop will help participants hone their own research plans, illustrate exemplary use of the data, and build community and connections between data users.

Prospective workshop participants are expected to have some familiarity with SAS, SPSS, or STATA. Enrollment is limited, with graduate students and recent Ph.D.'s particularly encouraged to apply. Tuition is $250, which includes course materials and lunches; a limited number of tuition wavers and travel scholarships will be available. The main instructor will be Trent Alexander, and other contributors will include Steven Ruggles, Miriam King, Carolyn Liebler (Assistant Professor of Sociology), Evan Roberts (North Atlantic Population Project Coordinator), and Matt Sobek.

While conferences, symposia, and workshops are designed to serve a clientele outside the University of Minnesota, other efforts serve an intra-university audience. We offer an annual graduate short course through the Humphrey Institute of Public Affairs on using IPUMS-USA and IPUMS-International data. In summer 2004, Alexander and Roberts offered a short course on "Global Perspectives on Population History" (based on IPUMS-International data) through the University of Minnesota's Institute for Global Studies. This course is scheduled to be repeated in Fall, 2006.

Data Use

Between the first beta release of IPUMS-International in Spring 2002 and the end of February 2006, 1,126 persons applied for access to IPUMS-International data. Senior project staff at MPC reviewed all applications and approved 718 applicants (64 percent of the total). Reasons for rejecting applications varied. Some proposed research plans were ill-suited to the project; for example, many required firm-level or genealogical data or information from countries not yet included in the database. Some unsuccessful proposals would have benefited profitmaking enterprises, rather than serving pedagogical or scholarly ends. Finally, research descriptions were sometimes insufficiently clear or detailed; in such cases, applicants were given the opportunity to reapply, contingent on providing more complete information.

About three-quarters (76 percent) of approved users are located in the United States. These U.S. users represent 150 research institutions. The remaining users are drawn from 43 other countries and 186 other institutions. Countries whose census data were included in the IPUMS-International database contribute a substantial share of the non-U.S. user pool. For example, Colombia, France, Brazil, Mexico, Kenya, and China were among the top twelve countries in terms of approved data users. Each new release that adds countries attracts researchers, educators, and students who previously had little or no access to their own nation's census microdata.

Users represent the following disciplines: Economics (40 percent); Demography (26 percent); Sociology (11 percent); Public Policy (6 percent); History (4 percent); and other fields, such as Geography, Public Health, and Political Science (13 percent). In addition to research application, the data are being widely used in classes on the following topics: labor economics; applied economics; econometrics; economic history; demographic history; demographic methods; migration; international development; statistics; quantitative methods in political science; quantitative methods in public policy; research methods in sociology; urban planning; and demographic data for policy analysis. In some cases, instructors built structured exercises around the data; in other cases, they encouraged students to conduct independent research using these data for class papers.

Though the majority of IPUMS-International users are housed at universities, researchers at other institutions also relied on the data. Particularly prominent are staff at international organizations such as the International Monetary Fund, the World Health Organization, the International Labour Office, the World Bank, the Inter-American Development Fund, and the United Nations. Registered users also work for national statistical agencies (e.g., from Brazil, Canada, China, Colombia, France, Greece, Kenya, and the U.S. Census Bureau)

and for national government agencies (e.g., the Chinese Academy of Sciences, the Kenyan Ministry of Health, the Mexican Ministry for Social Development, the Ugandan Ministry of Education, and the U.S. National Institute on Aging). Representatives from non-governmental research organizations and consortia (e.g., National Bureau of Economic Research, World Agroforestry Center, International Poverty Center, International Institute for Applied Systems Analysis, Center for U.S.-Mexican Studies, and Center for Global Development) are also using the data.

IPUMS-International census microdata offer great opportunities to study social and economic processes across time and space. While forty percent of users to date focus on a single country—addressing regional differences and/or change over time within national boundaries—the majority of users have adopted a cross-national comparative perspective. Specifically, 21 percent of users are studying two countries; 10 percent are studying three countries; 10 percent are studying four or five countries; and 19 percent are studying six to eight countries. Researchers have: made broad cross-national comparisons; compared a phenomenon in two similar countries; traced cohorts over time; checked the representativeness of survey data against census results; analyzed immigration via sending and receiving countries' data; developed demographic profiles of countries and population subgroups; empirically tested existing theoretical models; developed new models and methods; and indirectly estimated fertility and mortality.

The first non-beta version of the IPUMS-International database was released to the public in March 2003. The data have been available to scholars for only three years; yet the project is already beginning to yield publications and other research products. Specifically, we are aware (at the end of February, 2006) of 3 books, 4 book chapters, 20 journal articles, 3 dissertations, 3 working papers, 20 conference papers, and 13 courses based on IPUMS-International data.[6] The international scope and disciplinary range of journals publishing articles based on IPUMS-I data are striking. These journals include publications aimed at public health researchers (*Human Resources for Health, International Journal for Equity in Health*), historians (*Historical Methods, History of the Family, Continuity and Change, Hispanic American Historical Review, Revista de Indias*), computer scientists (*Statistics and Computing*), and demographers and sociologists (*Journal of Marriage and the Family, Scandinavian Population Studies, Sociological Methods and Research, Population Studies, Demography,* and *American Sociological Review).*

[6] These numbers are almost certainly an underestimates. Only publications reported to the MPC staff were at risk of being counted (1,799 items in all). Only a subset of these works—those whose title, abstract, or on-line text clearly and unequivocally referred to use of IPUMS-International data as opposed to IPUMS-USA data—were counted as scholarly products of the IPUMS-International for this report.

APPENDIX A
CURRENT STATUS OF DATA ACQUISITION AND PROCESSING

1. Summary table

Status	Number of countries	Number of samples
Agreements finalized		
A. Fully processed	8	28
B. May 2006 release	5	19
C. Other samples received	38	93
D. Other signed agreements	16	49
Subtotal (net countries)	63	189
Under negotiation	30	84
Total (net countries)	92	273

Note: China, France, Colombia, and Peru appear in multiple categories, but are only counted once in net totals

2. List of samples

A. Fully processed: 28 samples from 8 countries

Brazil	1970, 1980, 1991, 2000
China	1982
Colombia	1964, 1973, 1985, 1993
France	1962, 1968, 1975, 1982, 1990
Kenya	1989, 1999
Mexico	1960, 1970, 1990, 2000
United States	1960, 1970, 1980, 1990, 2000
Vietnam	1989, 1999

B. Scheduled for release May 2006: 19 samples from 5 countries

Chile	1960, 1970, 1982, 1992, 2002
Costa Rica	1963, 1973, 1984, 2000
Ecuador	1962, 1974, 1982, 1990, 2001
South Africa	1996, 2001
Venezuela	1971, 1981, 1990

C. Additional data received by MPC: 93 samples from 38 countries

Argentina	1970
Armenia	2000
Austria	1971, 1981, 1991, 2001
Belarus	1999
Bolivia	1976, 1992, 2001
Cambodia	1998
Canada	1971, 1981, 1991

C. Additional data received by MPC: 93 samples from 38 countries (continued)

Dominican Rep.	1960, 1970, 1981
El Salvador	1992
Egypt	1986, 1996
Fiji	1966, 1986, 1996
France	1999
Greece	1971, 1981, 1991, 2001
Guatemala	1973, 1981
Honduras	1961, 1974, 1988
Hungary	1970, 1980, 1990, 2001
Iraq	1997
Israel	1961, 1972, 1983, 1995
Italy	1981, 1991
Madagascar	1993
Malaysia	1970, 1980, 1991, 2000
Mongolia	2002
Netherlands	1960, 1970, 2001
Nicaragua	1971
Pakistan	1973, 1981, 1998
Palestine	1997
Panama	1960, 1970, 1980, 1990, 2000
Paraguay	1962, 1972, 1982, 1992, 2002
Peru	1993
Philippines	1960, 1970, 1980, 1990, 1995, 2000
Puerto Rico	1970, 1980, 1990, 2000
Romania	1992, 2002
Spain	1981, 1991, 2001
Sudan	1983, 1993
Turkmenistan	1995
Uganda	1991, 2002
United Kingdom	1991
Uruguay	1963, 1975, 1985, 1996

D. Signed agreements, data not yet received: 45 samples from 13 countries

Bangladesh	1981, 1991, 2001
Bulgaria	1985, 1992, 2001
Colombia	2006
Czech Republic	1970, 1980, 1991, 2001
Ethiopia	1984, 1994
Germany	1970, 1971, 1981, 1987, 1991, 2001
Indonesia	1971, 1980, 1985, 1990, 1995, 2000
Lesotho	1976, 1986, 1996
Malawi	1977, 1987, 1997
Mali	1976, 1987, 1998
Mozambique	1980, 1997

D. Signed agreements, data not yet received (continued)

Peru	2005
Portugal	1981, 1991, 2001
Slovenia	1981, 1991, 2001
Switzerland	1970, 1980, 1990, 2000

E. Under negotiation: 82 samples from 29 countries

Algeria	1966, 1977, 1987, 1998
Angola	1970, 1984
Burkina Faso	1985, 1996
Cameroon	1976, 1987
Chad	1993
China	1990, 2000
Congo, DR	1984
Finland	1960, 1970, 1975, 1980, 1985, 1990, 1995, 2000
Gambia, The	1973, 1983, 1993, 2003
Ghana	1984, 2000
India	1981, 1991, 2001
Ireland	1991, 2001
Jamaica	1982, 1991, 2001
Mauritania	1988, 2000
Mauritius	1983, 1990, 2000
Mexico	1995, 2005
Morocco	1982, 1994, 2004
Níger	1988, 2001
Nigeria	1991, 2006
Norway	1960, 1970, 1980, 1990, 2001
Poland	1978, 1988, 2002
Russia	1989, 1994, 2002
Rwanda	1991, 2002
Senegal	1976, 1988, 2002
Seychelles	1994, 2002
Sierra Leone	1985, 2004
Swaziland	1976, 1986, 1997
Tanzania	1988, 2002
Turkey	1970, 1975, 1980, 1985, 1990, 2000
Zambia	1980, 1990, 2000
Zimbabwe	1992, 2002

F. Countries that have declined to participate in the project

Australia	Latvia	New Zealand
Denmark	Lithuania	Norway
Estonia	Korea, Rep. of	Slovak Republic
Finland	Japan	Sweden
India	Jordan	Ukraine
Iran	Namibia	

APPENDIX C
TRIP REPORT, MINNESOTA POPULATION CENTER, 8/17/05

Jeff Naughton

Summary Recommendations and Comments:
The technical effort at the MPC is proceeding well. I have a couple of recommendations, some of which match those that the MPC folks have also made and are already working on:

1. As more and more data sets come online, there needs to be a more disciplined effort at providing good "provenance" information for each data release. By this I mean it should be possible to identify the processing programs that were used for any data set that anyone downloads from the MPC. This is much harder than it sounds because both the processing programs and the data products change frequently. The MPC staff mentioned going to a "release cycle" approach, I think that is a good idea.

2. Related to the previous point, as more and more data sets come online, each with a different but overlapping set of processing programs, managing these programs will become messy. I think there is no easy solution but approaches that will help include structuring the modules in the programs hierarchically (for example, multiple data sets require different versions of the "same" rule, structuring this as a base rule plus an inheritance hierarchy of specialized instances will help avoid chaos), frequent "regression testing" on subsets of data sets designed to stress the programs coupled with periodic full regression testing (to make sure "improvements" don't break things), and achieving a faster turnaround for the "propose data processing rule, see what it does, modify" cycle.

3. A side project investigating the utility and performance of an RDBMS in storing and extracting data sets would be helpful in evaluating the strengths, weaknesses, and future directions of the current approach to data extraction. This could be a good project for "outsourcing" as it doesn't require close involvement of MPC personnel.

Relationship to Computer Science Research
The MPC project is attacking a particularly messy instance of a problem that is attracting a lot of attention in research circles – the "data integration" problem, or, more completely, the "extract, transform, and load" problem. Advances in this area are slow and are (to date) always incremental – it just doesn't seem to be the kind of field where a few single "silver bullet" advances can further the state of the art dramatically. One place where the MPC might contribute to CS research would be to provide data sets to computer science researchers. The CS researchers could apply their techniques to these data sets and see where they do and do not match the solutions found by the MPC folks. It would also be a nice public example of the intricacies that can arise in data integration (such an example can be hard to come by since corporate data is so often shrouded in secrecy). Finally, even

the integrated data sets may be of some interest to test the performance of modern RDBMS technology against the fast inverted matrix system currently used by the MPC.

More Detailed Comments

Perhaps the most important thing for a technologist to understand is the nature of the problem the MPC is trying to solve. It is not a data storage problem; it is not really a data analysis problem; the real problem is how to take a bunch of wildly different data sets, each with their own internal inconsistencies and problems, and create an integrated data product that is easily understandable and accessible by social scientists. The team at the MPC is making good progress attacking this problem; in my opinion, the biggest challenges that they face arise because the fundamental problems they are encountering are hard, not because they are using inappropriate technology or approaches. They are moving to more uniform and standard data and program representations (storing things in XML, writing programs in Java with some visual basic.) One big and good move is moving to a "one version of the truth" setup in which there is one master copy of the metadata. This metadata captures issues things like what variables are available, where they came from in the source data, documentation about the variable both from a processing standpoint (what did the MPC people do with the variable, where did it come from on the original census form, etc.) This is a very positive move.

Probably the most important impression I had from the day was that the contributions of this project are the cleaned and integrated data sets but also all the intellectual hard work that went into their generation. Preserving all this intellectual hard work may be as important as the data itself, and the project should find ways to archive the data generation and transformation programs and the metadata that they generate. One way to think about this is that we don't want scientists 100 years from now who stumble onto this data to have to redo what the MPC people are doing now (and they will have to do a lot of it if all that is available to them is the final product data files.)

In the presentation the MPC folks arranged things in three main sections, I will follow that division here.

I. Data development.

By "data development" the team means the basic process of going from old source data to something that can be fed to later stages in their data transformation pipeline. At one level this is a very messy data integration problem. The source data includes data from all over the world gathered over the past century or so. Different censuses contain different data elements ("variables"), and figuring out exactly what they mean and mapping them to a single schema is a daunting task. This is exactly the problem that the whole "data integration" research area addresses. But in many ways this is an "AI-complete" problem at its heart and my guess is that it will defy total automation for a very long time if not forever. As one example, how would an automatic integration program decide to integrate two data sets in which one lists marital status as "married, single, divorced" and the other as "married, single serving in the military, single not serving in the military"? Or as another example, how would you integrate two census data sets in which one lists "dwelling type" as "house, apartment" and the other as

"thatched roof, tin roof, …"? (These examples are probably not real, but at the very least the project is facing equally difficult challenges.)

My point is that integrating such data sets will likely always require expert human intervention. The best you can do is provide tools that allow humans to investigate what a variable means and where it came from in various source data sets – that is, work on tools that make it easier for humans to understand and interpret the data rather than tools to automatically integrate the data.

One question the team asked me was whether their current approach of using Excel and Word files with visual basic routines to extract data elements was reasonable. I don't know of a better way that is as flexible, low cost, easy to replicate to mirrors, etc. as this approach. But I am far from an expert in this area so I could be wrong.

II. CDP – the Data Conversion Program

This refers to how the data is transformed to the "standard" IPUMS model once the mapping to this model has been agreed upon. Some of the transforms are very simple – e.g., converting from the numeric codes used by the source data to those used by IPUMS (e.g., converting from sex denotes by a 0 or 1 to sex denoted by a 1 or 2.) Other transforms are much more complex, especially those that fill in missing values or "correct" values that must be wrong. This is done by a complicated set of researcher-created rules that say things like "the 'mother' of the household can't be 10 years old" at the simple end to things like filling in missing values by identifying a set of predictor variables and using values from previous records that match the predictor variables.

Creating and debugging this rule set is difficult and time consuming. We discussed some techniques that might shorten the "propose rule, see what it produces, modify rule, try again" cycle. Pursuing these ideas will make the process more efficient but it will never be trivial.

One source of complexity is that the rules interact, and adding a rule to handle one case may "break" another. To catch such problems I think it would be good to adopt a discipline of "regression testing", where the new transformation rules are run and the results compared against the old ones, so that the IPUMS people rather than their users find the bugs. This could be done by running frequently running small data test data sets, and less frequently running the complete data set.

Another problem is organizing the reuse of the rules. Each data set has a different set of rules that are applied to it when it is processed. Some rules apply to multiple data sets. When a rule is modified to fix a problem in one data set, should the changed rule still be applied to every data set to which the original rule was applied? Or should the changed rule be a "new" rule that is only applied to the current data set? No one answer fits all. The most important thing here I think is to make sure that it is easy for the developer to answer questions like "where are all the places that this rule is used" and "does this change here for this data set break the rule for other data sets?" It seems like

these rules should be organized in hierarchies so that when one rule is split into two there is still a record of the relationship between the two.

A major issue here is that these programs constantly change, and for this reason the data sets that are generated change (actually it may be worse than that – even the data sets that are inputs to this process may change as data is corrected.) Moving forward it will be very important to be disciplined about all this. I think the project needs to move to a data release cycle, and needs to archive everything that went into the production of each data release. This will allow repeatability for people using the data set, and will help them understand if their results from an earlier release of the data don't match the results of a later release of the data.

III. Web site and Publishing

This part of the project appears to be going well. The web site is built using the "struts" web site development environment from Apache, and it is a huge improvement over the old "ball of Perl scripts and php" that used to be the system.

We spent some time discussing the extraction program. This is the program that actually "extracts" the data records and variables in response to user requests. It is a "query evaluation" system at its heart. Currently the requests from users are relatively simple and they are well served by the SDA tool from Berkeley. This is a "fast tabulation" system built on an inverted matrix model that is appropriate for the queries users currently ask. The MPC also has its own inverted matrix system that is somewhat simpler in that it requires fewer files to represent a data set.

There is a desire for users to ask more complex queries. This poses two challenges. The first is creating a user interface that lets them more easily compose their queries; the second is how a back end engine can evaluate those queries.

It is my suspicion that moving to an RDBMS will help here, but that is not entirely obvious. My guess is that the RDBMS will be substantially slower on the "bread and butter" queries that users currently ask than the fast tabulation tools. However, they will be able to answer much more complex queries, in fact queries so complex that would be hard for the users to express.

I think this is a good place for a small project to load one of the data sets in an RDBMS and then to see what happens. Questions to answer are "how does the performance of the RDBMS stack up"? "What new queries does it facilitate and are the useful? Is it realistic to ask users to pose these queries?"

APPENDIX D

INTERNATIONAL INTEGRATED MICRODATA SERIES: APPLICATION TO THE NSF HUMAN AND SOCIAL DYNAMICS COMPETITION

March 2004

Project Summary

A vast body of raw census microdata covering much of the world over the past four decades survives in machine-readable form. The bulk of these data, however, remains inaccessible to researchers. This proposal seeks funding to create an integrated global database of over 150 censuses from at least 44 countries. The International Integrated Public Use Microdata Series (IPUMS-International) will be the world's largest public-use population database, with multiple samples from each country enabling analyses across time and space. These microdata and accompanying documentation will be freely available for scholarly and educational research through a web-based data dissemination system.

The project leverages NSF investment in a major social science infrastructure project now nearing completion, "International Integrated Microdata Access System" (SBR9908380). That project covered many of the costs of finding and preserving microdata and documentation, negotiating dissemination agreements, developing data cleaning and sampling procedures, creating data conversion and dissemination software, and establishing design protocols for data and documentation. As a result, creation of the new database will be highly cost effective.

Intellectual merit. Census microdata represent an extraordinary untapped resource for research and education in human and social dynamics. With over five hundred million records spanning four decades, the new database will offer far broader chronological scope and greater sample densities than any alternative data source. For most countries, censuses are the most representative source of population data available. The new database will allow investigators to analyze global change during a period of unprecedented economic, demographic, and political upheaval.

The research team is uniquely qualified to undertake this massive database project. The investigators represent a diverse range of fields, including demography, economics, epidemiology, history, regional planning, and sociology. They have an unparalleled record of accomplishment on large-scale census infrastructure projects and wide-ranging international microdata experience. The Minnesota Population Center will house the project, providing exceptional facilities and technical support.

Broader Impact. This project will reduce barriers to international research and education by preserving datasets and making them freely available, converting them into a uniform format, providing comprehensive documentation, and implementing web-based tools for disseminating the microdata and documentation. The database will provide fundamental infrastructure for a broad range of fields in the social and behavioral sciences, including economics, geography, sociology, population studies, and environmental studies. Researchers in most countries do not presently have access even to their own national census microdata; IPUMS-International will democratize access to this vital scientific resource, creating unprecedented opportunities for global-scale research.

Most census data have traditionally been available only in aggregated tabular form. Census *microdata* provide information about individual persons, families, and households, and they allow users to interrelate any desired set of population and housing characteristics. The flexibility offered by microdata is essential for comparative research on social dynamics because the aggregate tabulations produced by national statistical offices are usually not comparable across time or between countries. In the few countries where census microdata covering multiple census years have been easily available to researchers, these data are the most widely used source for the study of large-scale economic and demographic transformations. Making integrated census microdata available for almost half of the world's population will allow researchers to describe the transformation of the world in far richer detail than previously possible. Even more important, these data will provide unprecedented opportunities to investigate the agents of change and assess their implications for human society.

Census microdata are an essential resource for studying large-scale transformational changes such as economic development, urbanization, fertility transition, large-scale migration, population aging, mass education, democratization, and growing international trade and capital flows. The availability of multiple censuses for each country lends historical depth, revealing the trajectories of change hidden in snapshots from the recent past. These data allow detailed study of the relationships of social and economic change to variations in climate, geography, and environment. They are also uniquely suited to assessing the human consequences of social, economic, and demographic transformations in such diverse areas as family structure, economic inequality, cultural diversity, and assimilation.

Objectives. This proposal seeks funding to develop critical infrastructure for the study of human and social dynamics. We propose to create the world's largest population database: an integrated census microdata series of unprecedented geographic and chronological scope. The database will become our most powerful resource for understanding the causes and consequences of the cataclysmic social and economic transformations that have reshaped the world during the past four decades. The project leverages NSF investment in a major social science infrastructure project now nearing completion, which has created a pilot version of the database, the International Integrated Public Use Microdata Series (IPUMS-International).

Most census data have traditionally been available only in aggregated tabular form. Census microdata provide information about individual persons, families, and households. Since the microdata include nearly all the detail originally recorded by census enumerations, users can interrelate any desired set of population and housing variables. The flexibility offered by microdata is particularly important for comparative research on social dynamics, because the aggregate tabulations produced by national statistical offices are usually not comparable across time or between countries. Unlike the basic social and economic aggregate indicators provided by the World Bank and other organizations, microdata allow researchers to conduct individual-level multivariate analysis and to design measures tailored to their particular research questions; unlike international surveys, census microdata offer sufficient cases for in-depth analysis.7

Although machine-readable census microdata exist for many countries, public access is restricted in virtually every case. We have negotiated license agreements with more than 50 countries—representing nearly half the world's population—to open access to these vital resources for scientific and educational purposes. This project will provide essential funding to exercise those agreements. But the goal of this project is not simply to make international microdata available; it will also make them usable. Even in the few cases where microdata are available, comparison across countries or over time is challenging due to inconsistencies between datasets and inadequate documentation of comparability problems. Because of these obstacles, analysts rarely attempt comparative international research based on pooled census microdata. This project will reduce the barriers to global-scale research by preserving datasets and making them freely available, converting them into a uniform format, providing comprehensive documentation, and implementing web-based tools for disseminating the microdata and documentation.

Specific aims. The tasks required for building the IPUMS-International database can be grouped into four major categories: (1) data preservation and democratization of data access; (2) data cleaning and processing; (3) documentation; and (4) dissemination.

---

[7] Much of what we know about global economic and social change derives from crude aggregate census statistics gathered from national statistical agencies by the United Nations, the U.S. Census Bureau, the World Bank, and similar organizations. Such data are typically national aggregates and provide few multivariate cross-classifications. Detailed tabular data are often available at the national level, but such statistics are rarely comparable across countries. International comparisons are also possible using individual-level survey data, such as the Demographic and Health Surveys. Such survey data, however, are of limited use for the study of large-scale human and social change; their small scale, limited chronological depth, and selective population coverage sharply restricts the potential methodological approaches and topics of investigation.

The work on data preservation, data processing, and documentation is often complicated by poor condition of the source materials. Most of the source files were designed for internal use by national statistical offices creating published census volumes. The data tapes—some of them over forty years old—are often difficult to read, and because of uneven documentation they are difficult to interpret. Computing was expensive and software was primitive in the 1960s and 1970s, so format errors often went uncorrected. Moreover, few statistical agencies devoted sufficient resources to documenting the internal files, so understanding them requires detective work.

Our first task is to continue our work on data preservation and democratization of data access. Many of the oldest files are already unreadable, and any delays in preservation will diminish the world's statistical heritage. Although we have negotiated agreements with over 50 countries to preserve and disseminate data, an enormous body of data remains at risk. We propose to extend our preservation efforts, with the goal of rescuing census files representing most of the world's population.

Once we have transferred the data to modern media, the second task is processing. Data processing includes standardizing data format and correcting format errors; assessing data quality and coverage problems; drawing high-density samples; identifying and correcting internal inconsistencies using logical and probabilistic procedures; allocating missing values; analyzing confidentiality risks and applying statistical confidentiality protections; and harmonizing variables.

Our third task—developing comprehensive documentation—is the most challenging and the most critical facet of our work. The central goal is to provide guidance to users on the meaning of census responses and their comparability across time and space. Since most of the files have minimal documentation, we will work from the original enumeration instructions and questionnaires in close consultation with experts from each country.

The final task is dissemination. We will distribute data and metadata (machine-processable encoded electronic documentation) through an integrated web-based data access system. This system will provide advanced tools for navigating documentation, defining datasets, constructing customized variables, and adding contextual information.

Results of prior NSF support. This project builds on thirteen years of NSF-funded experience integrating large census microdata samples. We began the Integrated Public Use Microdata Series for the United States (IPUMS-USA) in 1991 (SES-9118299, SBR-9422805). The IPUMS-USA project sought to harmonize, document, and disseminate all U.S. census microdata. In 1995, we began developing automated web-based tools for access to the massive body of data and documentation resulting from this effort (SBR-9617820). In 1999, we proposed extending the IPUMS paradigm to international data (SBR-9908380). The International Integrated Microdata Access System (IPUMS-International), now nearing completion, laid the groundwork for the present project and demonstrated its feasibility. In 2003, Ruggles received the Robert J. Lapham award from the Population Association of America in recognition of the contribution of IPUMS-USA and IPUMS-International to the field of demography. The following sections describe the results of these projects.

IPUMS-USA. External reviewers have described IPUMS-USA in glowing terms. Perlmann (2003), writing in the *Journal of American History*, praised the database as "one of the great archival projects of the past two decades . . . magnificent datasets." A reviewer for the National Institutes of Health described the IPUMS as "the crown jewels of historical demography in the United States." The review continued,

> Already, the availability of this unique sequence of historical evidence to scholars everywhere through efficient, user-friendly world-wide web access has begun to transform the face of

> historical and demographic research, and with it, our very understanding of the social context of our lives and the historical trajectory it is following. . . . The entire operation is a model for constructing the empirical foundation so vital to all research, the equivalent for historical demography of the human genome project (Center for Scientific Review 1999).

The IPUMS-USA database has quickly become one of the most widely used demographic resources in the United States. Since May 1996, over ten thousand researchers have registered to use the IPUMS data access system. These users represent an extraordinary range of academic disciplines; in addition to economists, sociologists, demographers, historians, and geographers, the user list includes researchers in such diverse fields as anthropology, epidemiology, environmental studies, political science, psychology, statistics, and even neuroscience.

The quantity of data accessed from our website is staggering. We are now distributing about 450 gigabytes of data per month, or an average of 638 megabytes per hour, 24 hours a day. We have prepared approximately 75,000 custom extracts of IPUMS-USA data and are now processing 2,000 data extract requests per month. This massive data distribution has borne fruit: at this writing, our bibliography lists approximately 1,000 books, articles, dissertations, and research reports (http://ipums.org/usa/research.php).

IPUMS-USA has generated a remarkable variety of research projects. Among the most studied areas are economic development, poverty and inequality, industrial and occupational structure, household and family composition, the household economy, female labor force participation, employment patterns, population growth, urbanization, internal migration, immigration, nuptiality, fertility, and education. Most IPUMS studies assess change over time, and IPUMS-USA has dramatically increased the number of quantitative studies of social and economic change in the United States.

Use of IPUMS-USA is not confined to academic research. The data are widely used by government agencies at all levels and by nongovernmental organizations for policy-relevant research. We have also received dozens of communications describing the successful use of the data series in graduate and undergraduate courses, and IPUMS-USA has even been used in primary and secondary classrooms. One of our private-sector partners, Key Curriculum Press, developed the Fathom software package to teach statistical literacy to young people. Using a simple interface, Fathom directly queries the IPUMS server for data extracts and allows children to perform basic statistical analyses, including significance testing. News media also rely on IPUMS-USA. At this writing, 120 journalists representing 50 newspapers and wire services have registered to use the IPUMS; 17 IPUMS-based articles have appeared in the *New York Times* alone.

IPUMS-International. In 1999, we applied to the NSF program "Enhancing Infrastructure in the Social and Behavioral Sciences" for a project entitled "International Integrated Microdata Access System" (10/01/99-9/30/04). Our goals were ambitious: (1) to inventory the world's surviving census microdata and preserve it wherever possible; (2) to obtain permission from national statistical agencies to redistribute the data; (3) to create an integrated database consisting of 80 censuses and 400 million observations; and (4) to develop a web-based system for accessing both the microdata and the metadata describing it.

The proposal requested the maximum funding level and project duration possible under the NSF infrastructure program guidelines: $10 million over ten years. The proposal reviewers were excited by the project's potential, but regarded it as high risk. Accordingly, we were awarded a smaller sum, $3.5 million over five years, to demonstrate the feasibility of the concept. NSF program officers assured us that if we could show success on a smaller scale within that period, the agency would look favorably on a proposal to finish the job.

The success of the pilot project has exceeded all expectations. We created a comprehensive inventory of known microdata, described in our award-winning book *Handbook of International Historical Microdata* (Hall, McCaa, and Thorvaldsen 2000), and preserved microdata from over 100 censuses. In May 2002, we released preliminary harmonized census microdata samples for Colombia, France, Kenya, Mexico, the United States, and Vietnam, followed by China in March 2003. By the end of the project in October 2004, we will finalize these samples, add five censuses from Brazil, and upgrade the data access tool and documentation. With samples from 29 censuses in eight countries, the database will comprise 152 million individuals residing in 40 million households.

This database—known as the International Integrated Public Use Microdata Series (IPUMS-International)—is freely distributed through a web-based data access system (http://ipums.org). In a very brief period, IPUMS-International has become an indispensable component of social science infrastructure. Hundreds of projects by scholars in 28 countries are already underway. In addition to university-based researchers, the user list includes representatives of many national statistical offices and international agencies such as the World Health Organization, the International Labour Office, and the World Bank. Research topics include the changing living arrangements of the aged, female labor-force participation and educational attainment, regional inequality differentials, patterns of age hypergamy, international migration, effects of emigration on labor markets, and relationships between divorce and family composition, between disease factors and education, and between educational attainment and cohort size. Most of these studies incorporate both cross-national and cross-temporal comparisons. For example, a National Academy of Sciences panel on "Transitions to Adulthood in Developing Countries" is using data from Colombia, Kenya, Mexico, and Vietnam to analyze changing outcomes such as schooling, work, fertility, and marriage as a function of age, gender, and household characteristics.

Despite this intensive research activity, IPUMS-International has limitations. Funding was provided to create samples for just a scattering of countries around the globe, so the database represents a fraction of the world's population. To capitalize on the potential of international census microdata for illuminating the global agents of social and economic change, the database must be greatly expanded.

The new project builds on this groundwork and encompasses four major activities. First, we will obtain data for approximately 85 additional censuses to ensure preservation for future generations. Second, we will dramatically expand the size and geographic scope of the integrated database. Third, we will improve the documentation for the entire database, exploiting new XML metadata standards to deliver codebooks and ancillary documentation tailored to the needs of particular research projects. Finally, we will upgrade the IPUMS data access system to incorporate powerful new tools, many of which we envisioned in the original $10-million version of IPUMS-International.

Strengths of the IPUMS model. Why are IPUMS-USA and IPUMS-International so successful? Key factors include the following:

- IPUMS data are available free of charge to researchers and educators with Internet access.
- IPUMS documentation is more comprehensive and more comprehensible than the original documentation produced by national statistical agencies.
- IPUMS provides a simpler but more powerful data extraction system than does any other provider of large-scale data samples. Remarkably, the IPUMS data extraction system is still the only system available that performs such basic and universal chores as rectangularizing hierarchical data files or concatenating multiple samples.

- IPUMS makes the coding schemes of each sample compatible with all other samples while simultaneously retaining all meaningful detail. This greatly reduces the labor involved in analyses using multiple census years and countries.

Most important is the power of the underlying data. The census files incorporated in both IPUMS-USA and IPUMS-International have four key strengths: broad chronological scope, large sample size, national coverage, and high data quality. Social scientists have increasingly recognized that we cannot understand contemporary social behavior without investigating processes of change. Many have turned to longitudinal sample survey data, which are invaluable for the study of short-run life-course transitions but unsuitable for the analysis of longer-term change across periods or between cohorts. In most countries, the census is the only source of microdata for the study of such long-run changes, and the IPUMS design makes such investigations comparatively simple. The second strength of the public-use census files is their large size and high sample density. The number of observations available for each census year ranges from the hundreds of thousands to the tens of millions. This allows the study of small population subgroups, such as ethnic minorities, specific geographic areas, or particular occupational groups. The large size of the census files also permits multilevel analyses of the effects of local conditions on individual and family behavior. The third strength, national coverage, is important because it allows researchers flexibility and permits generalization at the national level. Other datasets, such as employment surveys that cover only urban areas, have more limited applications (Vásquez, McCaa, and Gutiérrez 2001). Finally, censuses offer precision, reliability and response rates that compare favorably with the best alternative sources.

For most of the world, census microdata are either unavailable or restricted and are therefore seldom used.[8] In the United States, however, census microdata have been available to researchers for almost forty years and are the most widely used quantitative source for studying large-scale economic, social, and demographic transformations. For example, during the past two decades, census microdata have been the most frequently used source in the pages of *Demography*, the leading journal of population research. Even though the United States has abundant survey data, demographic researchers use census microdata three times as often as the next most popular data source.

Expected Significance. Between 1960 and 2000, the world's population doubled. Sharp interregional differences in growth rates—together with unprecedented urbanization and migration—led to dramatic spatial redistribution of population. Economic changes were equally remarkable. World per-capita gross domestic product roughly doubled, but that expansion was highly uneven, marked by growing inequality in many regions and little convergence in economic development between rich and poor countries. The most encouraging developments were in health and education; worldwide, child mortality dropped 60%, and adult literacy almost doubled. The extraordinary levels of demographic and economic growth also had ominous consequences: alarming environmental degradation and global climate change (World Bank 2002; O'Neill, Mackellar, and Lutz 2001).

The rate of population growth in the second half of the twentieth century was unprecedented and is unlikely to recur. In virtually every country, fertility rates are declining. This is creating another massive structural change: a shift in the age composition of the world's population, which will strain social resources as twentieth-century birth cohorts enter old age (National Research Council 2001). Other dramatic demographic trends—rising urbanization and international migration, industrialization

[8] Within the past decade, the Data Liberation Initiative in Canada and the Samples of Anonymized Records in the United Kingdom have made census microdata for those countries available to Canadian and UK researchers. Both initiatives sparked an explosion of new research, with hundreds of publications and thousands of users (e.g., http://www.ccsr.ac.uk/sars/publications/jointpub.htm).

of the developing world, and improvements in education and health—are likely to continue or accelerate in coming decades.

We know little more than the broad outlines of these global changes; our understanding has been hampered by the dearth of internationally comparable individual-level data. Making integrated census microdata available for half of the world's population will allow scholars to describe what happened in rich detail. Even more important, these data will provide unprecedented opportunities to investigate the agents of change and to assess their implications for human society. Census microdata are indispensable for studying large-scale transformational changes such as economic development, democratization, urbanization, fertility transition, expanding migration, population aging, growth of mass education, and expansion of international trade and capital flows. They are also uniquely suited for assessing the consequences of social, economic, and demographic transformations in such diverse areas as family structure, economic inequality, and cultural diversity and assimilation.

Research potential for IPUMS-International. The National Research Council (2001) recently made a compelling case for cross-national and cross-temporal data sources, declaring that "national and international funding agencies should establish mechanisms that facilitate the harmonization of data collected in different countries." The report argues that "cross national studies conducted within a framework of comparable measurement can be a substantially more useful tool for policy analysis than studies of single countries." The Council also recommended that "the scientific community, broadly construed, should have widespread and unconstrained access to the data." Scientific advances and policy insights are greatest when users with varying theoretical perspectives and models have access to the same data. IPUMS-International directly addresses these needs, by harmonizing microdata and metadata from a broad range of countries and distributing them to the research community through the Internet.

This proposed global microdata archive will be a permanent and substantial contribution to the infrastructure of human and social dynamics. By making these data easily accessible to researchers and developing comprehensive and comprehensible documentation, the project will stimulate new research that transcends national boundaries and static interpretation. Old census data are not of purely historical interest; rather, they are essential tools for basic social research and policy analysis. Models and descriptions of the past underlie both theories of social change and projections into the future. The IPUMS-International samples provide a unique laboratory for studying economic and demographic processes and for testing social and economic models.

The following are some of the most obvious topics for investigation.

Population aging. The policy challenges posed by population aging demand reliable projections and thorough analysis. New methods for projecting population aging require multi-dimensional parameters that must be derived from large microdata samples (Ahlburg, Lutz, and Vaupel 1999; Lutz and Goujon 2001; Inagaki and Matsuda 2003). Moreover, in many countries, census microdata are the only source with sufficient cases to study the aged population in depth (Palloni 2002).

International migration and assimilation. The censuses include questions on both migration and national origin. The data also provide good measures of intermarriage and other indicators of ethnic assimilation. IPUMS-International is one of the few data sources covering both sending and receiving areas. Analysts of immigration to the United States and Canada, for example, will be able to compare the characteristics of newcomers with those they left behind. Students of African, Asian, or Hispanic diasporas can compare the characteristics of people with the same ethnic origin in a wide variety of countries and assess how those characteristics have changed over time in each country.

Fertility decline. From 1960 to 2001, the world's total fertility rate declined from an average of well above 6 children to 2.8. This rapid transition has elicited much scholarly attention (Bongaarts 2003), but a great deal more remains to be done. IPUMS-International permits the study of differential fertility by occupational class, region, education, and a host of other variables at the individual, family, or community level. The richness of these data will illuminate the determinants of fertility decline in developing countries.

Environmental change. Used in combination with geospatial indicators on land-use patterns, climate change, endangered species, pollution, deforestation, desertification, and other environmental data, IPUMS-International provides unprecedented opportunities for multilevel analysis of relationships between social and economic transformation and changes in the physical environment.

Public health. The censuses collected a wide range of information relevant to public health, such as the availability of sanitation services, source of water supplies, type of cooking fuel, and housing construction material (De Vos and Arias 1996). Coupled with responses to questions on child survival and mortality, these data offer exceptional opportunities for pinpointing the correlates of public health at the local, regional, and national levels.

Comparative policy analysis. The availability of comparable microdata for dozens of countries with wide variation in public policies opens opportunities for natural experiments assessing policy outcomes. In the United States, this strategy has enables scholars to evaluate the effects of state-level variations in public assistance programs, access to health care, and tax policy (e.g., Duncan and Hoffman 1992; Lundberg and Plotnik 1995; Moffitt 1992; Ruggles 1997; Whittington 1993). IPUMS-International will allow similar fixed-effects models applied across countries to assess the impact of policy changes on economic development, inequality, urbanization, and demographic change.

These topics are only representative examples of the extraordinary research opportunities created by the integrated microdata series. Other key areas of investigation include the demography of violence, social correlates of physical disabilities, changes in household and family composition, urbanization, internal migration, work of women and children, union formation and dissolution, education and the spread of public schooling, and transformation of occupational structures. The expanded IPUMS-International—spanning four decades of social, demographic, and economic upheaval—will comprise our single most important resource for studying the agents of change in human society.

We should also note the extraordinary potential of the data for geographic analysis. Even in the United States, where microdata are abundant, most research in human geography relies on aggregate census data. For all its strengths, the U.S. microdata include minimal geographic identification and limited sample density. IPUMS-International, however, identifies geographic areas containing as few as 20,000 persons and covers between 5% and 20% of the population in most countries, opening powerful new avenues for geographic and multilevel analysis. The present proposal does not include a cartographic component, since our immediate goal is to maximize the geographic and chronological coverage of the database, but the potential for this kind of research is clear.

In a very brief period, IPUMS-USA multiplied many-fold the volume of quantitative research on long-run change in the United States. We anticipate that the expanded IPUMS-International will have even more profound consequences for research in economics, sociology, population studies, geography, history, and political science. In addition to fostering scholarly investigations spanning time and space, the new database will contribute to social science education, bringing the excitement of discovery into graduate, undergraduate, and even high school classrooms.

Methods and Procedures. The project consists of four interrelated work components. The *data liberation* component will expand our efforts to rescue endangered census microdata around the world and to open scientific access to these data wherever possible. The *processing* component will convert the data into modern format, impose disclosure protections, create samples, correct errors, and harmonize datasets over time and space. The *documentation* component will develop XML-based metadata for the entire database, with particular focus on comparability issues. Finally, the *dissemination* component will develop powerful web-based software tools that will maximize the usefulness of this extraordinary resource. Each component depends on the others, and the timing of work must be closely synchronized.

Page limitations preclude detailed discussion of the methods and procedures we have developed for each of these components, but we have published extensively on these issues (e.g., McCaa and Botev 2003; McCaa and Thomas 2003; Ruggles et al. 2003a, 2003b; Esteve and Sobek 2003; Block and Thomas 2003; McCaa et al. 2003; McCaa and Ruggles 2002a, 2002b; McCaa and Ruggles 2000; Hall et al. 1999; Gardner, Sobek and Ruggles 1999; Sobek and Ruggles 1999; Ruggles, Sobek and Gardner 1996; Ruggles, Hacker and Sobek 1995; Ruggles 1995; Sobek 1995). After a brief description of our source data, the following sections depict our general approach in broad strokes. We then discuss data preservation, sustainability, work plan, and evaluation.

Source data. Over fifty countries around the world have agreed to a perpetual license for dissemination of census microdata dating from 1960 through 2003, and we expect several additional countries to approve the license agreement soon. Our current funding is insufficient to execute all these agreements and thereby ensure data preservation and access; the present proposal requests critical funds for this purpose.

The original IPUMS-International grant provided sufficient funding to process and release data from 29 censuses representing eight countries. We have obtained supplemental funding from the National Institutes of Health (NIH) adequate to acquire and process data from at least 20 Latin American and European countries (HD44154 and HD47283). With funding from the current proposal, we will be able to expand the database further, to encompass over 150 censuses from 44 countries representing almost half of the world's population.

The cost of fully processing every dataset we plan to preserve would exceed HSD funding guidelines, so we must be selective. Because the NIH supplemental funding focuses on Latin America and Europe, the data-processing component of this project will emphasize key censuses from Africa, the Middle East, and Asia. We have not yet finalized the list of specific countries and censuses for full processing and dissemination; we will select countries based on data quality, population size, availability of key variables, chronological depth, cooperation of local experts, and quality of documentation. At this writing, our top candidates are Algeria, Canada, Egypt, Germany, Hong Kong, India, Indonesia, Israel, Madagascar, Malawi, Malaysia, Mongolia, Morocco, Pakistan, Palestinian territory, Philippines, Russia, South Africa, Sudan, Thailand, Tajikistan, Uganda, and Zambia.[9] These censuses include a wide range of questions covering basic demographic characteristics, housing

[9] Space does not permit us to present information on the status of each of the 82 countries we are working with. At this writing, 36 countries have signed formal dissemination agreements; 12 countries have provided letters of intent; 8 countries have provided verbal agreements, and will, we expect, sign the formal agreement shortly; and we are still negotiating with 25 countries (in addition, the United States requires no agreement). In addition, we have preserved endangered microdata from 116 censuses and basic documentation (e.g., enumeration forms) from over 900. For information on the current status of each country, see http://ipums.org/international/status.htm.

conditions, occupation and labor force activity, fertility, education, family relationships, migration, and disabilities.[10]

Data liberation. The data dissemination agreements represent a sea change in the policies of national statistical offices. Thanks to the tireless efforts of Professor McCaa, we have achieved unprecedented success in securing uniform perpetual dissemination agreements; in the past, most statistical agencies refused to make census microdata freely available for scientific inquiry and education. These agreements represent a historic opportunity for research on human and social dynamics. By democratizing access, the agreements reduce the cost of research and greatly enhance opportunities for both national and cross-national studies.

Copies of the license agreements are attached to this proposal. Under the terms of the standard agreement, the national statistical authorities retain copyright to the microdata but cede authority to the Minnesota Population Center and other authorized distributors to disseminate the data in perpetuity to researchers who assent to an electronic confidentiality contract (see clauses 2-3). As detailed in our discussion of confidentiality protection below, the end-user is obliged to use the data solely for scholarly research and education, respect respondent confidentiality, prevent unauthorized access to the data, and cite the data appropriately. The Minnesota Population Center is obliged to share the integrated data and documentation with the national statistical agencies and to police compliance by users. The signed agreements are highly general and uniform across countries; details specific to each country such as fees and sample densities are negotiated separately with each national agency. Under a carefully planned legal arrangement, the Regents of the University of Minnesota are responsible for enforcing the terms of these accords. Any disputes with national statistical agencies will be settled by arbitration under the authority of the International Court of Arbitration in Paris.

The agreements take force on the payment of a one-time fee paid by the project, averaging $5,100 per census. The fees partially compensate the statistical agencies for the costs of recovering data, translating documentation into English, and providing technical support and staff consultation. The actual costs of these activities usually exceed the cost of the license fees, so the willingness of the agencies to provide the samples for such nominal fees represents a substantial contribution by the statistical agencies to the project.

The original NSF IPUMS-International project financed the extensive negotiations needed to secure these agreements, but our negotiating success outstripped our resources; the limited funds available allowed us to execute agreements with only 14 countries. The first action of this new project will be to execute the agreements we have already negotiated and thereby secure access to the microdata in perpetuity. Over the course of the next five years, we will continue to negotiate further agreements, to meet the goal of democratizing access to machine-readable census samples for most of the world.

Urgent action is needed to preserve this unparalleled source of knowledge for future generations. Despite our considerable accomplishments in preserving census data, much is still at risk of destruction from physical deterioration of storage media and from the attrition of the technical staff who can locate and interpret the original files. Accordingly, data liberation is our highest priority.

Confidentiality protection. The protection of respondent confidentiality is also of paramount importance. We disseminate microdata under strict confidentiality controls approved by each national statistical office. Two strategies safeguard confidentiality: statistical disclosure protections and

[10] The subject coverage of each census is detailed at http://ipums.org/international/coverage.htm.

confidentiality agreements. Our goal is to minimize disclosure risk without compromising scientific use of the data.

We begin by assessing disclosure risks posed by each dataset. Based on this analysis, we design technical safeguards to prevent the identification of respondents. In all cases, names and detailed geographic information are suppressed. Other procedures include the following:

- Swapping an undisclosed fraction of records from one administrative district to another to make positive identification of individuals impossible.
- Randomizing the sequence of households within districts to disguise the order of enumeration.
- Combining codes that reveal sensitive characteristics or identify very small population subgroups (e.g., grouping together small ethnic categories).
- Top coding, bottom coding, and rounding continuous variables to prevent identification.

Institutional safeguards supplement the statistical methods of disclosure control. Before obtaining data, individual researchers must complete an application for data access and sign an electronic license agreement (http://www.ipums.org/cgi-bin/ipumsi/ipumsireg.cgi). As part of the agreement, researchers must agree to do the following:

- Maintain the confidentiality of persons, households, and other entities. Attempts to ascertain the identity of persons or households from the microdata are prohibited, as are allegations that a person or household has been identified.
- Implement security measures to prevent unauthorized access to census microdata. Under our agreements with collaborating agencies, redistribution of the data to third parties is prohibited.
- Use the microdata exclusively for scholarly research and education. Researchers are not permitted to use the microdata for any commercial or income-generating venture.
- Report all publications based on these data to the Minnesota Population Center (MPC), which will in turn pass the information on to the relevant national statistical agencies.

In addition, researchers must propose a research project that demonstrates a scientific need for the microdata. Each application for access is evaluated by senior staff. Once an application is approved, the user password is activated, allowing controlled access to data. Penalties for violating the license include revocation of the license, recall of all microdata acquired, filing a motion of censure to the appropriate professional organizations, and civil prosecution under the relevant national or international statutes. Employees of the MPC who work with the census microdata also sign agreements to respect data confidentiality.

The safety record for public-use census microdata is apparently perfect, with no verified breach of confidentiality in four decades of use. Our procedures are designed to extend this flawless record.

Sample design and variance estimation. In most cases, our source data consist of the complete internal census microdata files originally used to create published census volumes. In some cases, we draw samples from the full files; in others, the national statistical agency draws the sample to our specifications. Sample density is negotiated separately with each country and usually falls between 5% and 20%.

The census samples must balance the competing goals of sample precision and maximum information for research. The samples are clustered by household because many analytic topics (e.g., family structure) require information about multiple individuals within the same unit. This clustering significantly increases the variance associated with most individual-level variables; the number of independent observations in each census file is the number of households, not the number of individuals. Standard errors in cluster samples depend on both the size of the sampled clusters and on

the homogeneity of variables within clusters (Kish 1992; Hansen, Hurwitz, and Madow 1953; Ruggles 1995; Korn and Graubard 1995, 1999). In the worst case, with perfect homogeneity within clusters, the standard errors for variables would be inversely proportional to the square root of the number of clusters rather than the number of individuals. Thus, race and poverty status, which are comparatively homogeneous within households, will have underestimated standard errors if clustering is ignored. Conversely, for variables that are heterogeneous within clusters, such as age and sex, household clustering has little effect on sample precision.

We will partially counterbalance the loss of efficiency resulting from clustered design by drawing stratified samples. Stratification reduces standard errors for both characteristics that are explicitly stratified and characteristics closely correlated with the stratification variables. Our specific stratification design varies across censuses because of differences in variable availability and the salience of particular characteristics. For example, race or ethnicity is a critical factor in some populations, but is less significant (and sometimes unmeasured) in more ethnically homogeneous countries. For most censuses, we will stratify by place of residence, socioeconomic status, ethnicity, household size, and household type. Before finalizing the design of each sample, we will test the design through replicate techniques to ensure that it yields adequate sample precision.

The sample designs used for census microdata yield variances that differ from a simple random sample of the same size. Nonetheless, researchers using census microdata almost universally apply methods designed for simple random samples. The resulting p-values and confidence intervals are inaccurate, and this can lead to erroneous research conclusions and policy recommendations.

In recent years, social science researchers have devoted increasing attention to the problem of variance estimation for complex sample designs, and the leading statistical packages now incorporate Taylor-series linearization procedures that simplify accurate variance estimation (SAS 1999; Stata 2001; SPSS 2003). Researchers using census microdata rarely take advantage of these tools, in part because the data sources seldom include sufficient information on stratification and clustering to apply the methods correctly. To address this problem, we will construct the variables necessary to exploit variance estimation software and develop user-friendly documentation and recommendations for variance estimation using each statistical package, with examples of typical analyses.

Because statistical tests are at the core of quantitative research in the social sciences, reliable variance estimation is vital; otherwise, the cumulative process of scientific research rests on a foundation of unstable inferences. Tools developed by the IPUMS-International project to help researchers perform valid statistical tests based on accurate variance estimates therefore represent a significant contribution to the scientific community.

Reformatting and correction of format errors. Systematic reformatting and cleaning of each dataset will involve analyzing the record structure, reformatting the data into a standard hierarchical format, applying internal consistency checks, and correcting data errors. Our experience has taught us that the oldest datasets—those dating from the 1960s and 1970s—generally pose the greatest problems, a consequence of the computing and data storage constraints of the time. Even the most recent samples, however, require effort to verify that they are free of data format problems (Esteve and Sobek 2003).

The raw data files are preserved in a variety of formats, and each format poses potential problems. *Rectangular* files are the simplest, with geographic, dwelling, household, and family information replicated on each person record. In *hierarchical* files, the microdata have as many as six nested record types describing each geographic area, dwelling, household, and individual; irregularity in the sequence of record types can create widespread data problems. *Linked* censuses are organized into multiple record types stored in separate files with common case identification numbers. Small

imperfections in the identification numbers can cause significant problems. Finally, *inverted matrix* samples store each variable in a separate file. This data structure is optimized for rapid tabulation, and it depends on a perfect sequence of cases within each file.

We begin by reformatting each sample into a simple hierarchical format consisting of a household record followed by person records for each individual in the household. Any geographic or dwelling-level information is replicated on each household record. This reformatting often exposes problems that cannot be identified from a detailed examination of data frequencies or cross-tabulations. Thus, the process of restructuring the data is an integral aspect of diagnosis and cleaning.

Due to space constraints, we cannot present our solutions to the wide variety of data format problems we have already encountered. Each census is different. We employ whatever internal data are available to develop a strategy for logical or probabilistic correction of errors. As we grapple with dozens of censuses of varying vintages, we will encounter new problems and develop new solutions.

Assessment of data quality, data editing, and missing data allocation. Before processing each census, we evaluate underenumeration estimates derived from post-enumeration surveys and demographic analyses carried out by statistical agencies or academic researchers (e.g. Feeney and Alam 2003). We then compare the microdata with published census totals to ensure completeness, and execute a battery of internal consistency tests to ensure data integrity. In light of the high cost of full census processing, we must focus our limited resources on censuses that meet minimum quality standards. The results of our assessments will be included in the IPUMS-International documentation.

Most of the censuses—even those with the highest data quality—have never been cleaned to meet the standards necessary for public-use data sets. We check for such things as households with no householder or multiple householders, householders with multiple wives in countries without polygamy, implausibly large households, and duplicate records. We also look for inconsistencies between household and person records, in the relationships among the persons in a household, and among the characteristics of individuals. For example, we check for clear contradictions between age and labor force status, marital status, educational attainment, and school attendance. Where data errors can be unambiguously identified, we flag them as inconsistent. Once the consistency checks are completed, we edit missing and inconsistent values. For example, if sex is missing, it is edited by logical inference from the family relationship field or based on the sex of a spouse. We will adapt our existing data editing software to the new samples, with adjustments as needed to accommodate cultural variations. We will identify all edits with an appropriate data-quality flag.

When missing or inconsistent items cannot be resolved through logical inference, we turn to probabilistic allocation procedures (Ford 1983). For each variable, we identify criteria for matching a donor record used to impute the missing or inconsistent value. We determine these criteria through analysis of the best predictors for each variable, and they can vary from census to census. For example, if school attendance is missing, then one might allocate the school attendance of the most proximate individual in the file who shares the same age, sex, ethnicity, and parental socioeconomic status. If a perfectly matched donor record cannot be found, the record that meets the largest number of criteria is used. The donated value is then subjected to consistency checks and is rejected if unsuitable. We will fully document all allocation and editing procedures. Allocating missing and inconsistent data increases the reliability of sample estimates and makes the samples easier to use. These procedures, however, are not suitable for all analyses. We therefore identify allocated data items with a data quality flag, and our data access system will allow users to recode altered cases to missing values automatically.

Harmonization. All variables in the international census samples are numerically coded, and the classifications are inconsistent across census years and countries. Reconciling these codes is a major part of this project. Because classification systems and variable design affect research strategies, we must develop our plans with care.

United Nations organizations have twice sponsored large-scale projects for regional harmonization of census microdata. The first, carried out by the United Nations' Centro Latinoamericano de Demografía (CELADE), included 29 Latin American censuses taken between 1960 and 1976 (McCaa and Jaspers-Faijer 2000). The resulting data series included only the lowest common denominator of variables available across all countries. About half the variables available in the original censuses were discarded, and much critical detail on such variables as occupation and ethnicity was eliminated from the harmonized version of the datasets. In the second United Nations harmonization effort, the Population Activities Unit (PAU) in Geneva created semi-compatible samples of the aged population in 15 recent European censuses (Botev 2000). The PAU project did not standardize coding schemes for complex categorical variables, such as religion, family relationship, occupation, ethnicity, or language. Only the simplest variables, such as age, sex, marital status, and employment status, were recoded into a common scheme. Thus, the PAU samples lost less detail than the CELADE samples, but international comparisons remain cumbersome.

Our design strategy is more ambitious than those of CELADE and PAU. Unlike CELADE, we retain all the detail provided in the original samples, except for the confidentiality suppression described earlier. Unlike PAU, we provide a truly integrated database, in which identical categories in different census samples always receive identical codes. We employ several strategies to achieve these competing goals. For simple variables, such as age and sex, the original variables are compatible, and recoding them into a common classification is straightforward. For more 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. Finally, trailing digits provide detail only rarely available. Future versions of our data access system will guide researchers to the level of detail appropriate for the particular cross-national or cross-temporal comparisons they are making.

The classification scheme for marital status illustrates the approach. Under the IPUMS-International design, the first digit of marital status has four categories: single, married/in union, separated/divorced/spouse absent, and widowed. This is the maximum number of categories consistently distinguishable across all samples in the database. The distinction between divorced and separated is not maintained in all samples, so these categories are combined in the fully comparable first digit of marital status. At the second digit, divorced and separated persons can be distinguished, as can formal marriages from consensual unions. The third and final digit differentiates among types of marriages (civil, religious, polygamous) available for select countries only.

Geographic variables pose the greatest challenges. Because of cost constraints, we cannot fully harmonize the lowest level of geographic information available in each country over time. Where feasible, however, we will create consistent definitions of large metropolitan districts and provide maps of administrative districts identified in the microdata and other ancillary geographic information. More extensive geographic harmonization will require a separate research project.

Most data transformations used to create harmonized variables are simple recodes of one value into another. For each variable, we develop a data transformation matrix that provides information on the location of the original variable in each sample, each original category value, and each new

standardized category value. These matrices are maintained in a relational database. To maximize efficiency, recoding operations are carried out with a C program operating as a sequential batch process. In some cases, information from more than one variable in the original census is needed to construct a new compatible variable; for example, one might need information on both province and subdistrict to identify a metropolitan area. Data transformation matrices cannot handle such complex transformations; we therefore must resort to customized programming solutions.

Harmonization of hundreds of variables across 150 censuses entails over a million data transformations. Each transformation must be planned, executed, checked, rechecked, and documented. This work represents approximately a fifth of the effort required for this project.

Constructed variables. In addition to recoding variables to maximize comparability, we will carry out processing to enhance usability. Some procedures are straightforward, such as adding compatible variables on household serial number, person number within household, census date, country code, size of unit, and case weights. Others are more complicated; some examples follow.

Each census includes data on households and the relationships of individuals within households. To facilitate analysis, we create individual-level variables describing interrelationships among family members. The most important of these are three pointer variables that give the person number within the household of each individual's own mother, father, and spouse. These variables help researchers create measures of kin characteristics, fertility, marriage patterns, and family composition that are tailored to their specific research questions and analytic strategies. We will provide other fully compatible variables describing family and household characteristics at the individual and household level. These include family membership, family size, number of own children, number of own children under five years old, and age of eldest and youngest own children. We also plan to design new constructed variables to describe household and family composition in ways that reflect the diversity of family forms we encounter.

In addition to variables describing family composition and interrelationships, we will construct variables describing socioeconomic status. A substantial number of censuses provide no direct information on income, so occupation, education, and housing information are often the most important indicators of socioeconomic status. For IPUMS-USA, we provided two occupation-based measures of socioeconomic status—Duncan's Socioeconomic Index and an occupational income score—and researchers have used both measures extensively. We are investigating alternative occupation and housing-based socioeconomic indicators to assess their feasibility and appropriateness for the international samples (Hauser and Warren 1997; Sobek 1995, 1996, 1997; Treiman 1977; Nakao and Treas 1992; Ganzeboom and Treiman 1996, 2003).

Documentation. The creation of comprehensive integrated documentation is central to the project and is among its greatest challenges. Fortunately, we have access to a superb collection of raw materials for this purpose. We have already inventoried, catalogued, and scanned a wide range of documentation for over 900 censuses, most of which we acquired when the Statistical Division of the United Nations donated its historical archive of enumeration materials (MPC 2001).

We will provide harmonized English-language documentation on each of the samples included in the database. This documentation will cover census enumeration procedures and instructions; definitions of households, dwellings, group quarters, and other enumeration units; scanned images of original-language versions of the census questionnaires; post-enumeration processing; and other pertinent source documentation. At the variable level, we will provide detailed descriptions including wording of census questions, universe definitions, frequency distributions, and variable codes. Comparability

discussions will describe any deviations of particular censuses from the standard variable definition and will address differences over time and across countries.

The documentation will also include guidance for users on variance estimation, appropriate use of allocated cases, deviations of the microdata from published tabulations, and assessments of data quality and underenumeration. We will fully describe our data processing procedures, including confidentiality edits, reformatting, error correction, allocation, sample designs, and harmonization. In addition to verbal descriptions of data manipulation, we will publish all computer code and transformation matrices used to recode variables, construct new variables, and correct errors.

The data series will require the equivalent of thousands of pages of documentation. To manage this quantity of information, the web-based metadata access system will limit the scope of information to only those elements relevant to a given research project, as defined by the user. By constructing documentation pages dynamically, we can customize the documentation to the needs of individual users. Suppose, for example, a user selects censuses only for Egypt. In such a case, comparability discussions will cover only the specific censuses selected and customized tables will give marginal frequency distributions restricted to those datasets. When we incorporate over 150 samples into the database, this ability to filter out extraneous information will be critical. Our goal is to provide documentation that devotes attention to subtle problems of comparability without overwhelming users with information they do not need.

Machine-understandable metadata. As we develop documentation, we need to be cognizant of the costs of long-run maintenance and sustainability. The experience of the IPUMS-USA project is instructive. The IPUMS-USA documentation now consists of approximately 2,800 web pages. Most of these are static pages, but an increasing number are dynamic pages constructed automatically when users request them. This arrangement has many advantages, but it also creates three problems. First, because the documentation is system specific and hardware dependent, long-term preservation is a concern. Second, the continuous process of editing and correcting individual web pages creates serious issues of documentation version control. Finally, the system is difficult to maintain. When a variable is altered, changes must be made in multiple places, including data-definition files, tables that drive the documentation and data extraction programs, and static HTML documentation pages. Any discrepancies among these files can lead to system failure or user confusion.

We will address these problems by creating machine-understandable metadata using the Data Documentation Initiative metadata standard (DDI). The DDI is a non-proprietary, hardware independent, neutral standard that preserves the content and relational structure of the full documentation. This archival standard, developed by the national data archives of seven countries in collaboration with major data producers, was designed to reduce the costs of long-term preservation and access to data. The DDI will reduce the costs of system maintenance and decrease the potential for documentation errors. In a DDI codebook, each item is tagged with information about its meaning using the Extensible Markup Language (XML). A DDI codebook has a machine-understandable structure that allows for automated processing by data access software. Once our data and documentation access system are driven by DDI-compliant metadata, we can modify a variable by changing its specifications in a single location, and the software will propagate that change throughout the system. This approach will increase the flexibility of the data access system and simplify the addition of new data files and variables; it will also address concerns about the sustainability of data and metadata (Block and Thomas 2003).

Dissemination. Data sharing is central to the project; effective dissemination is essential if the data are to be widely used. Both data and documentation will be distributed through an integrated web-based data access system.

The IPUMS-USA data access system pioneered web-based dissemination of large-scale datasets and has served as a model for many other social science data dissemination efforts. IPUMS-International calls for robust second-generation data dissemination software for use across a wide range of datasets. Like the current software, this secure data extraction system will allow users to merge datasets, select variables, and define population subsets. The new system will also offer advanced tools for navigating documentation, defining datasets, constructing customized variables, and adding contextual information. The data extraction tool will incorporate documentation browsing and search functions so users have easy access to comprehensive documentation as they design their analyses.

We will also provide on-line data analysis. Our surveys of IPUMS users have found high demand for this capability. The system will use an analysis engine developed by the Computer-assisted Survey Methods Program at the University of California, Berkeley. Survey Documentation and Analysis (SDA) uses an inverted-matrix data format and data packing techniques to deliver frequency distributions, cross-tabulations, means, correlation matrices, and simple regression analyses (OLS, Logit, and Probit) for millions of observations in real time. The system also handles case selection and basic recoding. We expect that by 2008, we will be able to tabulate the entire data series in less than a minute. As part of this project, we will develop a customized web interface for the SDA system tailored to census microdata analysis. The availability of user-friendly on-line analysis will substantially broaden the audience for census microdata, and even sophisticated researchers will turn to the system for exploratory analyses.

As we adapt the IPUMS-International data access system to the additional censuses, we will make every effort to keep it user friendly. Given the far greater complexity of the new database, however, we must innovate to keep access easy. To take one example, the current system presents the available variables as a simple pick list. As the number of variables and datasets grows, this approach will become increasingly unworkable, so we will develop new tools for navigating the variable list. Users will be able to search the variable list according to keyword 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 sample in the series. We will provide simplified pick lists focusing on the most commonly requested variables, as determined through analysis of extract logs. If there are multiple variables for the same subject—such as the occupation and industry variables—we will provide brief usage discussions explaining when a given variable is the best choice, and when others would be better.

With each extract, users will be able to download fully customized documentation, including the relevant variable descriptions, comparability discussions, marginal frequencies, and enumeration instructions. In addition to documentation designed for humans, we will generate customized metadata designed to be read by computer software. Specifically, we will offer data definition files for the leading statistical analysis programs (SAS, SPSS, and Stata) tailored to each data extract. We will also create customized codebooks marked up according to the DDI metadata standard.

The extraction engine is designed to take full advantage of the hierarchical structure of census data. We offer researchers the option of rectangular or hierarchical output files and allow users to select households or families based on individual-level characteristics. Future versions of the data access system will add additional features that exploit the hierarchical structure of the data by automating the creation of new variables describing the characteristics of subfamilies, families, and households.

Users will also be able to replicate data extracts used in published studies. The ability to replicate existing studies is essential to the scientific enterprise; it provides our fundamental means of understanding, evaluating, and building upon past research. The current IPUMS-USA data access system allows users to replicate or modify their own past data extracts. In the new access system,

each extract will come with a recommended citation incorporating a unique number for the particular extract, and we will encourage scholars to cite that number in their published work. The data access system will replicate extracts based on the extract number. Thus, any authorized user will be able to create and download exact copies of datasets cited in published research.

Preservation and sustainability. We noted above a key innovation promoting long-run preservation and sustainability: DDI-encoded metadata will ensure that the microdata and documentation remain usable even if the technological environment shifts. We are also concerned about long-term maintenance of the database. The MPC is not a permanent data archive, and future funding is uncertain. Accordingly, our dissemination agreements specify that datasets can be archived and disseminated by the Inter-university Consortium for Political and Social Research (ICPSR). To ensure preservation, we will mirror the entire system at ICPSR and will maintain off-site backups of the data and documentation in both Minneapolis and Ann Arbor.

International partners. This project is a collaborative effort. Although our core staff has broad experience with U.S. and international microdata, a project of this intricacy demands close collaboration with experts from participating countries. The national statistical agencies of each participating country will supply complete raw microdata or samples of microdata; gather data and documentation for each census, which sometimes involves substantial research into unpublished agency documents; answer questions relating to enumeration instructions, descriptions of enumeration procedures, post-enumeration processing, and sample designs; and evaluate both documentation and data transformations carried out in Minnesota. When we encounter problems interpreting data or documentation that cannot be resolved by the statistical agencies, we will contract with expert consultants in each country to help us resolve them.

Work plan and evaluation. The complexity of this endeavor is substantial; accordingly, tightly integrated project management is essential. The co-investigators will work closely together, with weekly meetings and daily interaction. Although all senior staff will share responsibility for design issues, each will focus on a different aspect of project management. Ruggles will oversee the entire project and directly supervise the dissemination component; McCaa will have primary responsibility for data preservation and democratization of data access; Sobek will manage data processing; King will be responsible for documentation; Ahlburg, Assaad, and Levison will evaluate data quality and verify data processing and documentation to guarantee accuracy; and Davern and Oakes will work on sample designs and variance estimation. Additional information on project roles and responsibilities appears in the budget justification.

Because of significant investment from the previous projects, start-up costs will be minimal and work can begin immediately. We have a substantial body of documentation for all countries; we have microdata for many countries already in hand; and we have negotiated final agreements with most of the other participating countries. In addition, we have developed effective data cleaning and sampling procedures, written much of the needed data conversion and dissemination software, and designed the basic harmonization protocols for both data and documentation.

During the first project year, we will analyze the documentation for 150 censuses to identify unforeseen problems and to design compatible coding systems for key variables. By December 2005, we will produce a comprehensive plan for the design of the entire database. We have tentatively scheduled a meeting of our Advisory Board (described below) for April 2006 in Minneapolis, where we will undertake a detailed country-by-country analysis of the project design. Refinement and development of the data access and documentation systems will occur throughout the project. We plan to convert the dissemination software to the DDI metadata standard by the end of 2006, and to add the extract replication system and other advanced data access features by 2008.

Data and documentation processing will proceed simultaneously on parallel tracks. We plan to work on all censuses simultaneously, to ensure that the harmonized coding design can accommodate all variations that occur in the data. We will release a preliminary version of the database incorporating a subset of key variables and basic documentation in September 2007. Thereafter, we plan annual releases with an expanded variable set and enhanced documentation; we will release the final version with all variables and full documentation in December 2009.

We will evaluate our progress through three mechanisms. First, and most important, will be our success in meeting the ambitious schedule of benchmarks outlined above: planning and design in 2005; software and metadata in 2006 and 2008; and data processing and dissemination in 2007, 2008, and 2009. Second, we will establish an Advisory Board to evaluate our work, and we will include those evaluations in the annual progress reports. Although we have not finalized the membership of the board, we plan to invite at least two representatives of countries included in the project; two representatives of the United Nations Statistics Division, the World Bank or other international statistical organizations; and three social scientists with broad international research experience. At the end of each project year, the Board will meet with the investigators to evaluate progress and propose modifications and improvements. We will also use electronic communications to maintain close contact with the Board throughout the project period. Our third method of evaluation will be annual user surveys inviting researchers to provide feedback on the database; we will summarize the results of these surveys in our annual report.

References

Ahlburg, Dennis A., Wolfgang Lutz, and James W. Vaupel. 1999. Ways to Improve Population Forecasting: What Should be Done Differently in the Future? In *Frontiers of Population Forecasting,* edited by W. Lutz, J. W. Vaupel, and D. A. Ahlburg. Supplement to Vol. 24, 1998, *Population and Development Review*. New York: The Population Council, pp. 191-198.

Block, William, and Wendy Thomas. 2003. Implementing the Data Documentation Initiative at the Minnesota Population Center. *Historical Methods* 36: 97-101.

Bongaarts, John. 2003. *Completing the Fertility Transition in the Developing World: the Role of Educational Differences and Fertility Preferences.* New York: Population Council.

Botev, Nikolai. 2000. PAU Census Microdata Samples Project. In *Handbook of International Historical Microdata for Population Research,* edited by Patricia Kelly Hall, Robert McCaa and Gunnar Thorvaldsen. Minneapolis: Minnesota Population Center, pp. 303-17.

Center for Scientific Review. 1999. Review Group Report, Public Use Microdata Sample of the 1910 Census. Social Sciences, Nursing and Epidemiology Review Group 3, Center for Scientific Review, National Institutes of Health.

De Vos, Susan and Elizabeth Arias. 1996. Using Housing Items to Indicate Socioeconomic Status: Latin America. *Social Indicators Research* 38:53-80.

Duncan, Greg J., and Saul D. Hoffman. 1992. Welfare Benefits, Economic Opportunities, and Out-of-Wedlock Births among Black Teenage Girls. *Demography* 27:519-35.

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

Feeney, Griffith and Iqbal Alam. 2003. In *Fertility, Population Growth and Accuracy of Census Enumeration in Pakistan: 1961–1998*, edited by A.R. Kemal, Mohammed Irfan, and Naushin Mahmood. Islamabad: Pakistan Institute for Development Economics.

Ford, Barry L. 1983. An Overview of Hot-deck Procedures. In *Incomplete Data in Sample Surveys*, edited by W.G. Madow, I. Olkin, and D.B. Rubin. New York: Academic Press, pp.185-207.

Ganzeboom, Harry and Donald Treiman. 1996. Internationally Comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations. *Social Science Research* 25: 201-39.

Ganzeboom, Harry and Donald Treiman. 2003. Three Internationally Standardised Measures for Comparative Research on Occupational Status. In *Advances in Cross-National Comparison: A European Working Book for Demographic and Socio-Economic Variables*, edited by Jürgen H.P. Hoffmeyer-Zlotnik and Christof Wolf. New York: Kluwer Academic/Plenum Publishers.

Gardner, Todd, Matthew Sobek, and Steven Ruggles. 1999. IPUMS Data Extraction System. *Historical Methods* 32 : 119-124.

Hall, Patricia Kelly, Catherine Fitch, Margot Canaday, Lisa Ebeltoft-Kraske, Carrie Ronnander, and Kathleen M. Thomas. 1999. IPUMS Metadata: Documenting 150 Years of Census Microdata. *Historical Methods* 32 : 111-118.

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

Hansen, Morris, William Hurwitz, and William Madow. 1953. *Sample Survey Methods and Theory*. New York: Wiley.

Hauser, Robert M. and John Robert Warren. 1997. Socioeconomic Indexes of Occupational Status: A Review, Update, and Critique. In *Sociological Methodology*, edited by Adrian Raftery. Cambridge: Blackwell Publishers, pp. 177-298.

Inagaki, Sjichi and Yoshiro Matsuda. 2003. Population and Socio-Economic Structure Simulation Using Microdata. Invited Paper on Statistical Aspects of Projecting Populations, *54th International Statistical Institute*, Berlin.

Kish, Leslie. 1992. Weighting for Unequal P_i. *Journal of Official Statistics*. 8(2): 183-200.

Korn, Edward L. and Barry I. Graubard. 1995. Examples of Differing Weighted and Unweighted Estimates from a Sample Survey. *American Statistician*, 49(3): 291-295.

Korn, Edward L., and Barry I. Graubard. 1999. *Analysis of Health Surveys*. New York: Wiley.

Lloyd, Cynthia B. editor (2005). *Growing Up Global: The Changing Transitions to Adulthood in Developing Countries*. Washington, D.C., The National Academic Press.

Lundberg, Shelley, and Robert A. Plotnik. 1995. Adolescent Premarital Childbearing: Do Economic Incentives Matter? *Journal of Labor Economics* 13: 177-200.

Lutz, W. and A. Goujon. 2001. The World's Changing Human Capital Stock: Multi-state Population Projections by Educational Attainment. *Population and Development Review* 27(2): 323-339.

McCaa, Robert, and Nicolai Botev. 2003. Integrating European Census Microdata. *Eurostat Working Party on Demographic Statistics and Population and Housing Censuses*. Luxembourg.

McCaa, Robert, and Dirk J. Jaspers-Faijer. 2000. The Standardized Census Sample Operation (OMUECE) of Europe, 1959-1982 [1995]: a Project of the European Demographic Center (CELADE). In *Handbook of International Historical Microdata for Population Research*, edited by Patricia Kelly Hall, Robert McCaa, and Gunnar Thorvaldsen. Minneapolis: Minnesota Population Center, pp. 287-302.

McCaa, Robert and Steven Ruggles. 2000. IPUMS International. In Patricia Kelly Hall, Robert McCaa, and Gunnar Thorvaldsen, eds, *Handbook of International Historical Microdata for Population Research.* Minnesota Population Center, pp. 335-346.

McCaa, Robert and Steven Ruggles. 2002a. The Census in Global Perspective and the Coming Microdata Revolution. In *Nordic Demography: Trends and Differentials*, edited by J. Carling. Vol. 13, *Scandinavian Population Studies*. Oslo: Unipub/Nordic Demographic Society, pp. 7-30.

McCaa, Robert and Steven Ruggles. 2002b. Proyecto Col-IPUMS: Harmonizing the Census Microdata of Colombia, 1964-2003. In *Homologación de los microdatos censales colombianos. 1964-1993. Memorias del taller Col-Ipums*, edited by Fernan Vejarano and Robert McCaa. Bogotá: DANE.

McCaa, Robert, Steven Ruggles, Matthew Sobek, and Albert Esteve. 2003. IPUMS-International: A Restricted Access Web-Site Providing Anonymized, Integrated Census Microdata for Social Science and Policy Research. 54th International Statistical Institute, Invited Paper Meeting No. 38, Microdata Access, Berlin, Aug 13-20.

McCaa, Robert and Wendy Thomas. 2003. Archiving Census Documentation and Microdata: Preserving Memory, Increasing Stakeholders. *Notas de Población* 29: 289-305.

Minnesota Population Center (MPC). 2001. *World Census Questionnaires from the Archive of the United Nations Statistical Division.* First edition. http://www.ipums.umn.edu/international/enumeration_forms.shtml.

Moffitt, Robert. 1992. Incentive Effects of the U.S. Welfare System: A Review. *Journal of Economic Literature* 30: 1-61.

Nakao, Keiko and Judith Treas. 1992. The 1989 Socioeconomic Index of Occupations: Construction from the 1989 Occupational Prestige Scores. GSS Methodological Report No. 74. Chicago: National Opinion Research Center.

National Research Council. 2001. *Preparing for an Aging World: The Case for Cross-National Research*. Washington, D.C.: National Academy Press.

O'Neill, Brian C., F. Landis MacKellar, and Wolfgang Lutz. 2001. *Population and Climate Change.* Cambridge: Cambridge University Press.

Palloni, Albert. 2002. Living Arrangements of Older Persons. *Population Bulletin of the United Nations* Special Issue, 42/43: 54-110.

Perlmann, Joel. 2003. IPUMS. (Web Site Review) *Journal of American History* June 2003 (Vol. 90, No. 1) pp. 339-340.

Ruggles, Steven. 1995. Sample Designs and Sampling Errors in the Integrated Public Use Microdata Series. *Historical Methods* 28: 40-46.

Ruggles, Steven. 1997. The Effects of AFDC on American Family Structure, 1940-1990. *Journal of Family History* 22: 307-25.

Ruggles, Steven. 2000. Data User's Perspective on Confidentiality. *Of Significance . . . Journal of the Association of Public Data Users* 2:1-5.

Ruggles, Steven, J. David Hacker and Matthew Sobek. 1995. Order out of Chaos: The Integrated Public Use Microdata Series. *Historical Methods* 28: 33-39.

Ruggles, Steven, Miriam King, Deborah Levison, Robert McCaa, and Matthew Sobek. 2003a. IPUMS-International. *Historical Methods*, 36: 60-65.

Ruggles, Steven and Matthew Sobek, et. al. 1997. *Integrated Public Use Microdata Series: Version 2.0*. Minneapolis: Historical Census Projects, University of Minnesota.

Ruggles, Steven, Matthew Sobek, Miriam L. King, Carolyn Liebler, and Catherine Fitch. 2003b. IPUMS Redesign *Historical Methods* 36: 9-21.

Ruggles, Steven, Matthew Sobek and Todd Gardner. 1996. Distributing Large Historical Census Samples on the Internet. *History and Computing* 9: 145-159.

SAS. 1999. *Documentation for SAS Version 8*. Cary, NC: SAS Institute, Inc.

SPSS. 2003. *Correctly and Easily Compute Statistics for Complex Sampling*. Chicago, Illinois: SPSS Inc.

Stata. 2001. *Reference Manual.* College Station Texas: STATA Press.

Sobek, Matthew. 1995. The Comparability of Occupations and the Generation of Income Scores. *Historical Methods* 28: 47-51.

Sobek, Matthew. 1996. Work, Status and Income: Men in the American Occupational Structure Since the Nineteenth Century. *Social Science History* 20: 169-207.

Sobek, Matthew. 1997. A Century of Work: Gender, Labor Force Participation, and Occupational Attainment in the United States, 1880-1990. Ph.D. diss., University of Minnesota.

Sobek, Matthew and Steven Ruggles. 1999. The IPUMS Project: An Update. *Historical Methods*. 32: 102-110.

Treiman, Donald. 1977. *Occupational Prestige in Comparative Perspective*. New York: Academic Press.

Vásquez, Gabriela, Robert McCaa, and Rodolfo Gutiérrez. 2001. 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* 6: 151-78.

Whittington, Leslie A. 1993. State Income Tax Policy and Family Size: Fertility and the Dependency Exemption. *Public Finance Quarterly* 21: 378-98.

World Bank. 2002. *World Development Indicators 2002*. CD-ROM. Washington, DC.