

MINNESOTA POPULATION CENTER



UNIVERSITY OF MINNESOTA

Working Paper Series

**Using Cyber-Resources to Build Databases for Social
Science Research**

**Matthew Sobek, Monty Hindman, and Steven Ruggles
Minnesota Population Center
University of Minnesota**

June 2007

Working Paper No. 2007-01

Abstract

The Integrated Use Microdata Series (IPUMS) is the premiere infrastructure project supported through the NSF Human and Social Dynamics Priority Area. Over the next four years, the IPUMS-International project will release data and metadata from approximately 150 censuses of 45 countries, totaling about a half-billion records and some 20,000 variables. Because of the unprecedented scale of this work, we have had to develop innovative cyber infrastructure for both data processing and dissemination. The source data consists in most cases of raw microdata captured by census enumerations during the past 50 years, usually in obsolete formats with paper documentation. Our greatest challenge is development of comprehensive machine-processable encoded electronic documentation, or metadata. This metadata underlies every aspect of IPUMS-International data processing work, including standardizing data formats 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. The same metadata drives our integrated web-based data access system that provides advanced tools for navigating documentation, defining datasets, constructing customized variables, and adding contextual information, as well as a basic set of on-line data analysis tools.

Background

The IPUMS-International project began in 1999 with a social science infrastructure grant from the National Science Foundation, “Integrated International Microdata Access System” (NSF 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 confidentiality risks and applying protections, and harmonizing coding across countries and censuses. These data, together with the accompanying documentation covering temporal and international comparability issues, are freely available to researchers through a web-based data access system (<http://ipums.org/international>).

IPUMS-International is already an important data resource. Over 1,000 projects by scholars in 50 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 2004, we received substantial new funding to expand the database.¹ We have obtained and archived data covering over a billion person records from 185 censuses in 58 countries, and the datasets are still flowing in (see Appendix A). 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.

To make this vast body of data accessible, the pace of our work has accelerated dramatically. In our first six years—from 1999 to 2005—IPUMS-International released 28 census samples from eight countries. In 2006, we more than doubled the size of the database, adding 35 new samples and 12 new countries. Our plans call for releasing a similar body of 30-35 datasets every year for the next five years.

The massive increase in the scale of the project 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,

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.

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 paper describes how we are getting the job done.

Software and Metadata Infrastructure

This project has required us to develop a substantial body of new software and metadata.²

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 in 2005 and early 2006, but it is continuously refined and improved.

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.

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

² Design of these systems was carried out under the direction of Peter Clark, Monty Hindman, Catherine Ruggles, and Matt Sobek, and the software engineers were Marcus Peterson and Colin Davis. The design benefited greatly from the input of Jaideep Srivastava of the University of Minnesota Department of Computer Science and Jeffrey Naughton of the University of Wisconsin's Department of Computer Science, as well as Nupur Bhatnagar, a Computer Science graduate student.

inconsistent data items; and generates frequencies for each variable. We revised this software substantially in 2005 to operate on a new XML-based metadata structure.

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. In 2005, we replaced the PHP script initially used for IPUMS-International dissemination with a new Java-based system. Like the data conversion program, the new dissemination system operates on a new XML-based metadata structure. In addition, we replaced hundreds of pages of static HTML pages with dynamic documentation pages generated on the fly.

All the software for data preparation, data conversion, and dissemination is driven by metadata. Metadata is formally structured documentation of digital data. We have developed a comprehensive metadata system for IPUMS-International, with a goal of capturing everything we know about the data in a structured format that can be processed by machine. Our specification is in some respects similar to the Data Documentation Initiative (DDI) Document Type Definition developed by a consortium of data archives and producers, but it handles additional kinds of metadata required by our project.³ The IPUMS-International metadata format is compatible with DDI and we will be able to generate DDI-compliant codebooks for datasets on demand.

³ The DDI is described at <http://www.icpsr.umich.edu/DDI/>.

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, 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.

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>
```

For viewing and editing by humans, we display the information in tabular format without tags. We are presently 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 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 was 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 redesigned our work process to accommodate the necessary increase in data production. 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 was not scalable. 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 system, from the input materials provided by national statistical offices to the dissemination of the database on the web.

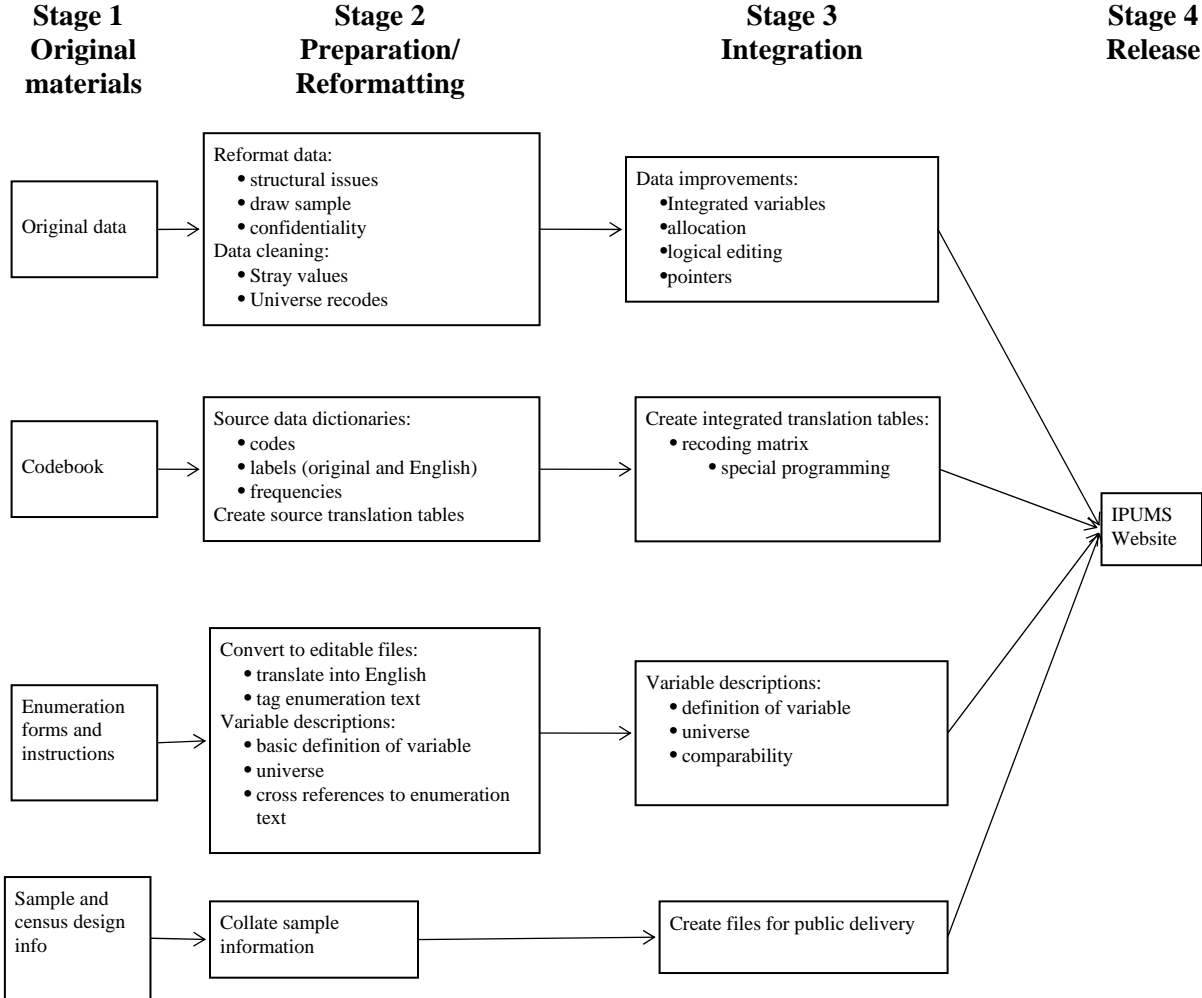
Figure 2. Part of Costa Rica 2000 Data Dictionary (untagged view)

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 dígitos			CR00A423
					1		Agriculture, cattle ranching, hunting, and		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	

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 the metadata is an essential aspect of data processing. 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 highly 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 of our revamped work process 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 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 variable. 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 possible.

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. To increase the scale of data production, however, we 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, we 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. 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 Zambia in 1996 and 2001, Belarus in 1999, and Spain in 1981 and. 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, and the number of cases for that category, in brackets.

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.

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 and insert the needed software code at the bottom of the table.

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. Beginning in December 2006, user are also able to examine the constituent source variables that served as input to the integrated version.

Figure 4. Part of IPUMS translation table for RELATE

code	label	za1996a	za2001a	by1999a	es1981a
rectype	RELATE	P	P	P	P
columns	Relationship to head of household				
norecode					
hide					
proj1		ZA96A406	ZA01A407	BY99A401	ES81A402
proj2					
proj3					
<tt>					
1000	HEAD	1 = Head of household {846410}	1 = Head/Acting head {947351}	1 = Person recorded first within house	1 = Principal person {579383}
1000	"				
1000	"				
1000	"				
2000	SPOUSE/PARTNER	2 = Husband/wife/partner {426080}	2 = Husband/wife/partner {408540}	2 = Wife, husband {217576}	2 = Spouse of principal person {45655}
2100	Spouse				
2100	"				
2100	"				
2200	Unmarried partner				
2200	"				
3000	CHILD	3 = Son/daughter {1407983}		3 = Daughter, son {311597}	3 = Child of principal person {889117}
3000	"				
3100	Biological child		3 = Son/daughter {1286439}		
3100	"				
3100	"				
3200	Adopted child		4 = Adopted son/daughter {38674}		
3200	"				
3300	Stepchild		5 = Stepson/stepdaughter {19316}		
3300	"				
3400	Child/child-in-law				
3500	Child/child-in-law/grandchild				
4000	OTHER RELATIVE				
4100	Grandchild	7 = Grandchild {408044}		9 = Granddaughter, grandson {31694}	5 = Grandchild of principal person {28}
4100	"				
4110	Grandchild or great grandchild		9 = Grandchild/great-grandchild {4850}		
4120	Great grandchild				
4130	Great-great grandchild				
4200	Parent/parent-in-law				6 = Father or mother of principal pers
4210	Parent	5 = Father/mother {43269}	7 = Parent {31514}	4 = Mother, father {8326}	
4210	"				
4210	"				
4211	Stepparent				
4220	Parent-in-law		8 = Parent-in-law {11722}	6 = Mother-in-law, father-in-law {3832}	
4300	Child-in-law		10 = Son/daughter-in-law {30795}	7 = Daughter (sister)-in-law, son (broth	4 = Son/daughter-in-law of principal pe
4300	"				
4400	Sibling/sibling-in-law				
4410	Sibling	4 = Brother/sister {124820}	6 = Brother/sister {142034}	5 = Sister, brother {3990}	
4410	"				
4420	Stepsibling				
4430	Sibling-in-law		11 = Brother/sister-in-law {27212}		

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++. The program operates from the same XML 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.

A relatively new feature of the DCP 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. Dissemination

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. The IPUMS website is based on Java software driven by our XML metadata. Most pages on the website are constructed on the fly, including the main variables page, the variable description and codes pages, and the data extraction system. If a user clicks on our main variables page or a variable description, for example, a Java program builds an HTML page from the metadata.

In 2006, we introduced several improvements to the dissemination system to enhance usability; the most important of these are described in the following paragraphs.

- *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 can view the relevant enumeration text pertaining to the underlying

census question(s) for any or all samples. They can also view any particular piece of enumeration text within the context of its parent document. Finally, images of the original documents are also available, allowing users to view the enumeration forms and instructions in the original language.

- *Metadata filtering.* Because the software constructs pages dynamically, we can filter the content that is put on the screen based on user-defined preferences. For example, users 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). As noted, virtually all of our web pages are 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.
- *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. The hierarchal nature of the coding schemes is now explicitly recognized by the IPUMS-International software; users can choose the simplified version of a variable and eschew the extra detail.

These enhancements allow researchers to plan analyses and design customized datasets in a rich informational environment.

Future enhancements. The substantial long-term investment in the IPUMS database is only justified if the data are widely used to produce important new discoveries about population and society; accordingly, investment in dissemination is essential. We plan to harness the power and expertise of the IPUMS user community by means of a suite of Web 2.0 tools. Approximately 15,000 active IPUMS users represent an extraordinary resource; by developing participatory software—such as wikis, analytical tool-sharing, and special interest research groups—we can efficiently leverage this resource, dramatically improving user support and training at modest cost.

APPENDIX A

CURRENT STATUS OF DATA ACQUISITION AND PROCESSING

1. Summary table

Status	Number of countries		Number of samples	
Agreements finalized				
A. Fully processed	20		63	
B. May 2007 release	6		17	
C. Other samples received	45		116	
D. Other signed agreements	15		35	
Subtotal (net countries)	73		221	
Under negotiation	27		75	
Total (net countries)	97		296	

Note: Some countries appear in multiple categories, but are only counted once in net totals.

2. List of samples

A. Fully processed: 63 samples from 20 countries

Belarus	1999
Brazil	1970, 1980, 1991, 2000
Cambodia	1998
Chile	1960, 1970, 1982, 1992, 2002
China	1982

Colombia	1964, 1973, 1985, 1993
Costa Rica	1963, 1973, 1984, 2000
Ecuador	1962, 1974, 1982, 1990, 2001
France	1962, 1968, 1975, 1982, 1990
Greece	1971, 1981, 1991, 2001
Kenya	1989, 1999
Mexico	1960, 1970, 1990, 2000
Philippines	1990, 1995, 2000
Romania	1992, 2002
South Africa	1996, 2001
Spain	1981, 1991, 2001
Uganda	1991, 2002
United States	1960, 1970, 1980, 1990, 2000
Venezuela	1971, 1981, 1990
Vietnam	1989, 1999

B. Data received by MPC: 116 samples from 45 countries

Argentina	1970, 1980, 1991, 2001
Armenia	2000
Austria	1971, 1981, 1991, 2001
Bolivia	1976, 1992, 2001
Canada	1971, 1981, 1991, 2001
China	1990
Czech Republic	1991, 2001
Dominican Rep.	1960, 1970, 1981
El Salvador	1992
Egypt	1986, 1996
Fiji	1966, 1986, 1996
France	1999
Guinea	1983, 1996
Guatemala	1964, 1973, 1981, 1994, 2002
Honduras	1961, 1974, 1988, 2001
Hungary	1970, 1980, 1990, 2001
Indonesia	1971, 1976, 1980, 1990, 1995
Iraq	1997
Israel	1961, 1972, 1983, 1995
Italy	1981, 1991
Madagascar	1993
Malawi	1987, 1998
Malaysia	1970, 1980, 1991, 2000
Mali	1987, 1998
Mauritius	1990, 2000
Mexico	1980, 1995, 2005
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
Portugal	1981, 1991, 2001
Puerto Rico	1970, 1980, 1990, 2000
Rwanda	1991, 2002
Saint Lucia	1982, 1991
Sudan	1983, 1993
Turkmenistan	1995
United Kingdom	1991, 2001
United States	2005
Uruguay	1963, 1975, 1985, 1996
Venezuela	1961, 2001

C. Signed agreements, data not yet received: 35 samples from 15 countries

Bangladesh	1981, 1991, 2001
Bulgaria	1985, 1992, 2001
Colombia	2005
Czech Republic	1970, 1980
Dominican Rep.	2002
Germany	1970, 1971, 1981, 1987, 1991, 2001
Indonesia	2000
Lesotho	1976, 1986, 1996, 2006
Malawi	1977
Mali	1976
Mozambique	1980, 1997
Nicaragua	1995, 2005
Peru	2005
Slovenia	1981, 1991, 2001
Switzerland	1970, 1980, 1990, 2000

D. Under negotiation: 75 samples from 27 countries

Burkina Faso	1985, 1996
Cameroon	1976, 1987
Chad	1993
China	2000
Ethiopia	1984, 1994
Finland	1960, 1970, 1975, 1980, 1985, 1990, 1995, 2000
Gambia	1973, 1983, 1993, 2003
Ghana	1984, 2000
India	1981, 1991, 2001
Ireland	1991, 2001
Jamaica	1982, 1991, 2001
Kenya	1969, 1979
Mauritania	1988, 2000
Morocco	1982, 1994, 2004

Niger	1988, 2001
Nigeria	1991, 2006
Norway	1960, 1970, 1980, 1990, 2001
Poland	1978, 1988, 2002
Russia	1989, 2002
Senegal	1976, 1988, 2002
Seychelles	1994, 2002
Sierra Leone	1985, 2004
Swaziland	1976, 1986, 1997
Tanzania	1988, 2002
Thailand	1970, 1980, 1990, 2000
Turkey	1980, 1985, 1990, 2000
Zambia	1980, 1990, 2000

E. Countries that have declined to participate in the project

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