IPUMS

Working Papers

Building Research Infrastructure for Harmonized International Census Microdata

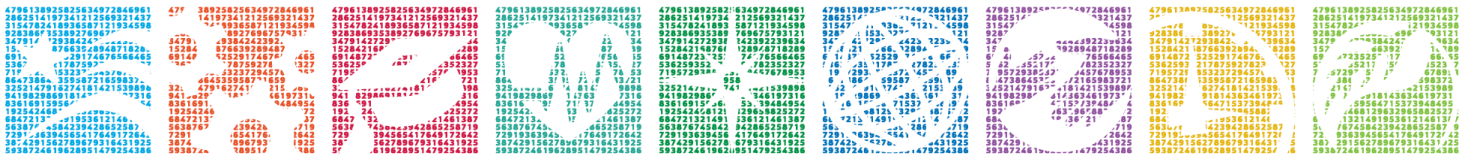
Matthew Sobek†
University of Minnesota

Lara Cleveland
University of Minnesota

September 2020

Working Paper No. 2020-01
DOI: https://doi.org/10.18128/IPUMS2020-01

†Address correspondence to Matthew Sobek: sobek@umn.edu. Support for this research was funded by IPUMS International: NIH R01HD044154, NIH R01HD0427283, NSF SES0851414.



Abstract

IPUMS International harmonizes and disseminates census microdata collected over multiple decades by roughly 100 countries. There is little commonality in the source material over time within countries and no consistency at all across countries. To manage this heterogeneity, IPUMS has developed an extensive data infrastructure driven by metadata. Researchers manage correspondence tables to assign disparate input codes into a common global classification scheme for each categorical variable. These self-documenting tables govern the data harmonization software and provide the value labels for the web dissemination system and generation of statistical package syntax. IPUMS does not transform data using code, unless complex logic is required. Stemming from a single metadata source, the data always stay in sync with the web dissemination system. Other metadata components describe the input data and govern the display of samples and variables in the web system. The development of infrastructure driven by metadata empowers the research staff who best understand the data to accomplish the vast majority of the tasks required for harmonization. Because of the complexity and scale of data harmonized across so many sources, a sophisticated dissemination system is essential, and is an integral part of our approach. Harmonized data are more complicated than data from a single source, and it is essential to convey information without overwhelming researchers.

Introduction

IPUMS International is the world's largest collection of population microdata available for research. It is largely composed of census data collected since the 1960s by the National Statistical Offices of dozens of participating countries. IPUMS harmonizes the census variables over time and across countries, so the same code has the same meaning in all times and places. The goal is to facilitate comparative international research. In pursuing this goal, IPUMS does not reduce international differences to a set of least common denominators, but aims to provide researchers access to the full detail of the original data. Many countries do not have a dissemination mechanism for their census microdata; thus, IPUMS serves as the sole practical means to access much of this data (Minnesota Population Center 2019).

Census data are the backbone of national statistical systems. Most countries incur great expense to conduct censuses—a testament to their perceived value. Censuses aim to enumerate entire national populations, giving them uniquely broad coverage. IPUMS has reached agreements with over 100 national statistical offices to disseminate their census microdata, conditioned on their use for research and teaching (Meier, McCaa, and Lam 2011; Ruggles et al 2015). As of 2020, the database included over 370 censuses, including a subset of historical datasets dating back to the 18th century. IPUMS has recently begun adding selected household surveys to the database as well, to provide more recent and intercensal data. Most countries contribute multiple censuses to the database, allowing the study of change over time. The post-1950 census data are samples, but they are larger than any survey. In contrast, many of the older census datasets from Europe and North America include every resident in their countries (Ruggles et al 2011). The median IPUMS census sample includes 10 percent of the national population and has 840,000 person records. Thirty-two samples exceed 10 million records. In total, the database includes individual-level information on over one billion persons. The large sample sizes allow the study of small subpopulations that surveys may have insufficient cases to analyze. The IPUMS samples are nationally

representative and typically offer geographic detail to the second administrative level within countries, such as counties, districts, or municipalities.

Censuses are very much a national enterprise and are products of their times. Consequently, the source data IPUMS receives are inconsistent across censuses within countries, and they are thoroughly incompatible across countries. Both the roster of variables and their classifications vary from one census to the next. Over the decades, statistical offices have subscribed, to varying degrees, to international standards for census question wording, but the standards themselves have evolved over time (United Nations 2017). Even similarly worded census questions do not necessarily yield consistently coded microdata due to processing decisions. The documentation of the data is sometimes fragmentary—especially for older data—and it is usually in the official national language. Many of the census microdata files were never edited for use outside the statistical office, being conceived primarily as a means to produce the published census reports for the period they were conducted. Consequently, it is not unusual to encounter inconsistencies among the responses for an individual, across individuals in a dwelling, or between the dwelling record and the person records associated with it. In general, the older the data, the more common such issues are.

The size and complexity of the database has forced IPUMS to innovate. We must contend with both heterogeneity of the source material and the large scale of the data, which totals well over a terabyte. Conventional approaches to data management and dissemination are not well suited to microdata of this size. Censuses typically ask fewer questions than surveys, but harmonization of variables from a hundred organizations with differing languages and statistical traditions is both logistically and conceptually difficult. Conveying the resulting complexity to researchers without overwhelming them is a significant dissemination challenge. Efficiently filtering information and subsetting the database is essential to make it usable. For this reason, we consider the data access system to be an integral component of our approach to harmonization (Sobek and Cleveland 2017).

A signature feature in the design of IPUMS has been the development of software driven by metadata (Sobek, Hindman, and Ruggles 2007). The earliest implementation of this approach was the use of correspondence tables to govern recoding of the original variables into harmonized classifications. We subsequently added a variety of other metadata types that document input data files, control web display, describe sample designs, and provide machine-actionable versions of census questionnaires. The technical infrastructure built around these metadata components is maintained by a team of software developers. The metadata empowers the research staff to perform the functions that affect the subject matter of the population data they understand without being burdened by technical complications best left to programmers. The metadata-driven system is also highly flexible, easing the addition of new samples to the database.

Data Pre-processing

Data arrive from partner countries in many different formats, and IPUMS must convert them into a consistent form suitable for further processing. IPUMS software is designed to run on fixed-column ASCII files in a hierarchical structure, with a household record followed by a person record for each resident. To achieve this format, sometimes multiple hierarchical record types must be collapsed, different file types might need to be merged, or dwelling-level information (such as geography) must be extracted from persons in a rectangular file to create household records. Some minimal language translation of labels is sometimes required to inform the reformatting stage.

IPUMS has developed utility programs to deal with the most commonly encountered formatting issues, but pre-processing still regularly requires sui generis programming solutions. The goal is to create a single consistent format that the rest of the IPUMS software infrastructure can act upon without downstream customization. Data formatting can also uncover issues in the original data—especially in the organization of person records into households. A variety of checks are necessary to ensure the soundness of that structure, which is critical for both technical and substantive research purposes. Problems might

include households without heads in a de jure census, households with multiple heads, fully blended households, straggler records, or other problems. These irregularities must be rectified by logical inference, explicit identification of household fragments, or whole-household record donation.

When a statistical office has an existing scientific-use census sample, those are usually the data they contribute to IPUMS. In cases where the statistical office draws a sample for the project, it is often done to IPUMS specifications. But roughly 30 percent of the time, the country supplies their full-count data and relies on IPUMS to draw a sample. These are usually instances of older censuses from developing countries that lack the resources to process files that may not have been touched for many years. The IPUMS sampling scheme is intentionally simple to explain and execute: a 1-in-N systematic sample of dwellings. Because of geographic sorting, every administrative unit receives proportional representation in the data, amounting to low-level geographic stratification (Cleveland, Davern, and Ruggles 2011). The original full-count data are archived for preservation, but only the sample is disseminated for research.

The final stage of pre-processing involves steps to ensure the data are sufficiently anonymized to prevent identification of individuals (McCaa, Ruggles, and Sobek 2011). None of the data IPUMS receives ever contain names, but they may have too much subject or geographic detail to ensure confidentiality. We suppress very small categories and top- and bottom-code thin tails of continuous variables. We also swap a small number of households across geographic units to add an additional degree of uncertainty to re-identification efforts. Finally, areas with fewer than 20,000 population in recent censuses are combined with adjacent units until they achieve that threshold. The user registration license also prohibits any attempt to identify individuals in the data.

Variable Standardization

After the data are consistently formatted, each file is subjected to a series of pre-harmonization steps to create fully specified input for the harmonization stage. This process centers on metadata development. A "data dictionary" is developed for each dataset, recording its layout and the

characteristics of its variables. These highly structured metadata files initially include basic variable-level information: column locations, variable labels, codes, and value labels. When labels are in a language other than English, they are translated, so all subsequent work can be in a common language. An automated tool collects frequencies from the associated data file and inserts them into the data dictionary, flagging any undocumented values. Because the data dictionaries are properly structured metadata, they are suitable subjects for various utility programs. For example, they can be used to generate SAS, SPSS, and Stata set-up files to read the input data while applying variable and category labels. These syntax files, used for internal analysis, are ephemeral products that can be regenerated at any time as the dictionaries evolve.

In the next step, we specify how the original variables will be converted into more standardized "source variables." We create a set of fields in the data dictionary paralleling those that document the codes and labels for the original variables. By editing the contents of these parallel fields, research staff can relabel and recode the original data while the original codes and labels are retained for future reference. The resulting unharmonized source variables become the input data for the subsequent harmonization stage.

The goal in creating source variables is not to recode the data into common classifications, but rather to rationalize and fully document each sample-specific variable. Early in the life of the project we determined that there was too much variation and were too many irregularities in the original data to resolve every issue while trying to harmonize the data cross-nationally. The problem needed to be broken into pieces; hence this intermediate stage. In developing the source variables, we consolidate stray values into designated missing value categories, label and numerically code blank values typically representing the not-applicable cases, and clarify the labels, which may be ambiguous or have been translated poorly from another language. At the end of this stage of processing, every categorical value has a label and has been numerically coded. Meaningful categories in continuous variables, such as the not-applicable value,

top-code, or missing values, are also labeled. A brief variable description is written to document each variable, in the process confirming that we understand its meaning. The sample-specific source variables are available to IPUMS users for downloading, alongside the globally harmonized variables that apply to all samples. By providing access to the source variables, we ensure that no meaningful detail in the original data is lost. The source variables free us from the need to harmonize rarely available variables, and allow us the option to suppress some detail during the harmonization process, if necessary, to keep a variable from becoming unwieldy.

Figure 1 shows a small part of a data dictionary. The leftmost fields indicate the five variables being displayed (sex, relationship, marital status, etc.). Some rows contain are variable-level information, while the intervening rows pertain to specific values. The "Original Variable" columns document the values in the data as provided to IPUMS. We translate the Spanish labels into English and collect frequencies, in the process identifying several undocumented values in RELATE and MARST. The "Standardized Variable" columns are used to convert the original variables into the source variables, cleaning up stray values and editing labels as necessary. For MARST, we use the "SCode" column to reassign the original undocumented values 7 through 9 to all receive the value "9" with a label of "unknown" in the source variable. The final column in the Figure documents the universe of people who should have a response for the variable (i.e., they were asked the census question from which the variable was derived). A number of other fields in the data dictionary describe additional features of each source variable, such as implied decimals, how or if it should be displayed in the web dissemination system, and other attributes.

Figure 1. Data dictionary

				ORIGINAL VARIABLE				STANDARIZED VARIABLE		
Name	Column	Width	Variable label	Code	Original label	Translated label	Frequency	SCode	Output label	Universe
SEX	**40**	**1**	**Sex**							**All persons**
				1	Varon	Male	131,612	1	Male	
				2	Hembra	Female	140,478	2	Female	
RELATE	**41**	**1**	**Relationship**							**All persons**
				1	Jefe	Head of household	48,508	1	Head of household	
				2	Conyuge (esposa)	Spouse	18,749	2	Spouse	
				3	Companera	Partner	15,989	3	Partner	
				4	Hijo	Child	138,276	4	Child	
				5	Conyuge del hijo	Spouse of child	883	5	Spouse of child	
				6	Companera del hijo	Partner of child	946	6	Partner of child	
				7	Otro parentesco	Other relative	29,278	7	Other relative	
				8	Domestica	Servant	2,717	8	Servant	
				9	No pariente	Not related	16,733	9	Not related	
				0		[no label]	11	99	Unknown	
MARST	**42**	**1**	**Marital status**							**All persons**
				1	Soltero (nunca casado)	Single, never married	217,960	1	Single, never married	
				2	Casado	Married	46,122	2	Married	
				3	Viudo	Widowed	5,019	3	Widowed	
				4	Divorciado	Divorced	1,665	4	Divorced	
				5	Separado legalmente	Legally separated	906	5	Legally separated	
				6	Anulado	Annulled	406	6	Annulled	
				7		[no label]	9	9	Unknown	
				8		[no label]	2	9	"	
				9		[no label]	1	9	"	
CONS	**43**	**1**	**Consensual union**							**Persons age 15+**
				1	Si	Yes	36,352	1	Yes	
				2	No	No	97,797	2	No	
				0	Menores de 15 anhos	Under age 15	137,941	9	NIU (not in universe)	
LIT	**44**	**1**	**Literacy**							**Persons age 5+**
				1	Si	Yes	141,514	1	Yes	
				2	No	No	79,384	2	No	
				3	No responde	Undeclared	6,787	8	Unknown	
				0	No aplica	Not applicable	44,405	9	NIU (not in universe)	

The data dictionaries are ultimately ingested into the IPUMS metadata database where they are accessed by data transformation and web software. One of the most important metadata elements, and one which requires considerable effort to specify, is the universe of respondents for each variable. We examine the census questionnaires and empirically verify the universe for each variable, because we find that statistical office processing sometimes alters the theoretical population at risk. Universes for constructed variables can be particularly ambiguous. In addition, the out-of-universe cases—often represented as blanks or zeroes in the original data—are sometimes combined with missing values or meaningful zeroes. We use programming to separate the out-of-universe and other cases as necessary. The empirical verification of universes has a practical side benefit: it forces the staff to engage the data in a multivariate way, which can reveal issues not evident from simple marginal frequencies.

Another important step in the development of the data dictionaries involves connecting each source variable to the specific questionnaire text that pertains to it. To make this possible, the original

census questionnaires and enumerator instructions are first converted into machine-actionable metadata from their original formats (typically pdfs). The image files are translated as needed into English and entered into a simple text document. XML tags are inserted in the file to provide basic formatting for web display, and every distinct block of text is assigned an ID number. We enter in the data dictionary the text block numbers associated with the census question(s) from which each source variable was derived. Using these tags, web software can compile questionnaire text on demand for users, but the tagging serves an important internal need as well. The questionnaire wording is the most fundamental documentation for most variables in the source data, and it is invaluable to easily access that wording during harmonization.

Data Harmonization

Data harmonization involves the development of variables that span countries and times. This requires determining which variables are conceptually the same across datasets. Those determinations cannot be made solely on the basis of variable names and labels, but may require reference to codes, value labels, census question wording, or even category frequencies. It can sometimes come down to a judgement call: weighing the value of user convenience against the possibility of misleading researchers by combining variables with differing shades of meaning or strikingly different population universes. Even where concepts appear equivalent, there may be a fundamental incompatibility in their classifications. For example, variables pertaining to counts may be grouped into incompatible intervals in different samples, or censuses may combine response items in overlapping ways that defy harmonization without extreme aggregation and loss of detail. If concepts or categories differ significantly, we create parallel harmonized variables to minimize the likelihood of user error.

The core activity of data harmonization is to equate variable codes and labels across samples, so each category means the same thing across all censuses (Esteve and Sobek 2003). The primary instrument for achieving this is a correspondence table ("translation table") like the one depicted in Figure 2 for the variable "Class of worker." The columns on the left show the harmonized output values and their labels.

Each column on the right represents an input dataset: in this case census samples from four countries spanning a thirty-year period: Ecuador, Romania, Venezuela, and Tanzania. Note that the full translation table for this IPUMS variable contains over 300 samples. Each row in the translation table contains items that are conceptually the same and that thus receive the same codes in the output data. The work is performed by a researcher using tools we have developed specifically for this process. In broad strokes, the process is as follows: a researcher identifies the source variables in the different samples, a program inserts the values and labels for those variables into the translation table from the appropriate data dictionaries, and a researcher then aligns the codes and assigns output codes and labels (the "harmonized codes" columns on the left). This sort of semantic integration is intellectual labor that no computer program can perform. The absence of a category in the input data can be as meaningful as the presence of one. The work requires a holistic view of the universe of codes for each sample and consideration of the underlying questionnaire text, especially for some of the more challenging variables (Sobek and Cleveland 2017).

Figure 2. Translation table: Class of worker

HARMONIZED CODES		INPUT CODES			
Code	Label	Ecuador 2001	Romania 2011	Venezuela 1981	Tanzania 2012
000	NIU (not in universe)	9 = Blank (N/A)	9 = Blank (N/A)	0 = Blank (N/A)	99 = Blank (N/A)
100	Self-employed		1 = Self-employed		
110	Employer	1 = Employer		7 = Owner with employees	1 = Employer
120	Working on own account	2 = Self-employed		8 = Own-account worker	
121	Own account, agriculture				4 = Own account, agriculture
122	Own account, non-agriculture				3 = Own account, non agric
200	Wage/salary worker		2 = Wage or salary worker		2 = Employee
210	Wage worker, private employer	5 = Private sector employee			
211	Non-manual worker, private			2 = Private sector professional	
212	Manual worker, private			4 = Private sector manual labor	
213	Domestic worker			6 = Domestic service	
220	Wage worker, government				
221	Federal, government employee	4 = State employee			
222	Local government employee	3 = Municpal employee			
223	Non-manual worker, govt			1 = Public sector professional	
224	Manual worker, government			3 = Public sector manual labor	
300	Unpaid worker		3 = Unpaid worker		
310	Unpaid family worker	6 = Family worker		5 = Unpaid family worker	5 = Contributing family worker
320	Apprentice				6 = Apprentice
400	Other		4 = Other		7 = Other not specified

Variable harmonization is designed to retain all the detail in the original samples while providing a fully integrated database in which identical categories in different samples always receive identical codes. We employ several strategies to achieve these competing goals. In cases where original variables are compatible and recoding is straightforward, we write documentation noting any subtle distinctions between samples. For some variables, it is impossible to construct a single uniform classification without losing information from samples that are detail-rich. 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.

The classification scheme for "class of worker" in Figure 2 illustrates the composite coding approach. The first digit of the variable identifies four substantive categories consistently available in all samples: 1) self-employed, 2) wage-salary worker, 3) unpaid worker, and 4) other. Take the “Self-employed” category as an example (codes 100-122). The Romania sample does not make the distinction between employers and own-account workers, thus these categories are combined at the fully comparable first digit, receiving a code of 100. Other samples distinguish between employers and working on own account, as reflected in the second digit of the code. The third and final digit differentiates among types of own-account workers (agriculture and non-agriculture). The one-digit and multi-digit versions of the composite variables can be accessed as their own distinct variables in the IPUMS database. For many researchers, the single-digit version is sufficiently detailed while minimizing comparability issues.

The translation tables exemplify our metadata-centered approach. We do not write recode statements, except in exceptional circumstances. We write software to read our metadata. Simply moving an item from one cell to another in the translation table accomplishes the recode. The benefits are significant: a researcher can readily interpret the coding decisions while seeing all the associated labels with their codes and frequencies. If a new code is needed to handle some variation introduced by a

sample, the researcher simply adds a row in the table and aligns the appropriate input codes to it. The translation tables also help with sustainability. Reorganizing the codes to accommodate a new sample is quite easy compared to sifting through a mass of impenetrable logical assignment statements. Thus our system is far less error-prone and is much more adaptable than what could be achieved in a statistical package or simplistic approach to data processing. IPUMS is a living project, and we can never know the full universe of labels and coding structures that will need to be incorporated into the existing harmonized variables in the future. The metadata-driven translation tables provide a practical solution to this challenge.

The custom IPUMS data conversion program reads the translation tables to produce the integrated output data. There are, however, instances where translation tables cannot accommodate the logic required to recode a variable, and conventional programming is required; for example, for recoding continuous numeric variables like income into categories or combining multiple input variables. The data conversion program supports modularized programming in which discrete variable-specific logic can be written by research staff without affecting the main structure of the application maintained by software developers.

Geography variables pose a unique harmonization challenge. The census samples typically report first- and second-level subnational geography for place of residence or previous residence. These are administrative divisions specific to the period in which the data were collected. These subnational units—especially the more detailed second level—can merge, split, or change boundaries between censuses. The goal of harmonization is to construct units that share the same exact spatial footprint across census years (Kugler, Manson, Donato 2017). The only way to do this is to combine units, because we do not usually have the detail in the microdata to disaggregate them. The creation of spatially harmonized variables requires obtaining GIS boundary files or digitizing old maps. The process involves overlaying the boundary files across censuses and combining units as needed until all boundary changes occur within the

aggregated units and no changes cross their borders. The resulting geographies are stable over time, so researchers studying change can be assured that their analyses are not an artifact of differing populations. The original unaltered units are retained in separate census-specific geography variables, for researchers who require maximum detail at a specific point in time.

Harmonized Documentation

Harmonized data inevitably are more complex than the original. Composite coding and clear labeling can go only so far in conveying the compromises involved in combining items derived from unique questionnaires processed by dozens of organizations. IPUMS is a general-purpose research tool, and it is impossible to predict which differences in the underlying data might be critical for a specific researcher's analysis. Our solution is to write harmonized variable documentation that highlights comparative issues to encourage exploration by the user. The documentation is intended as a component of an integrated data dissemination system. The IPUMS web system provides unified access to each variable's codes and frequencies, population universes for each sample, questionnaire text, and links to the underlying unharmonized source variables. The documentation we write does not seek to exhaustively explain all potential issues, which would produce too much detailed text for most users to cope with. The aim instead is to write enough to alert the reader to the issue and point them to the metadata element where they can explore further for themselves.

Variable documentation is initially drafted during data harmonization, noting any decisions or underlying differences that are not self-evident from the codes and labels. The comparability text notes differences over time within countries, and a general comparability section describes cross-national issues. The text might note, for example, when some additional detail is present in a source variable that could not be accommodated in the harmonized variable. Perhaps the reference period for the question differs between countries, or the wording of the question was unusual in one or more samples, possibly meriting examination of the questionnaire text. One of the most persistent comparability issues that

cannot be conveyed via codes and labels involves differences in the population universe for the question. For instance, if only persons age 15 and above are asked the employment questions in a census, that can distort comparisons of child labor rates in censuses that applied a lower minimum age. Significant universe differences always warrant mention in the variable description, but it is advisable for researchers to review that component of the metadata for all their key variables.

Data Processing

The IPUMS data conversion program (DCP) is a custom C++ application that uses the translation tables and other metadata to produce a globally harmonized output file corresponding to each input dataset. Because every record must be processed sequentially, there would be no advantage to using a database for this work, which would impose significant costs in terms of speed and overhead. Output data can be produced as ASCII or parquet-format—the latter promising more efficient dissemination options as we make changes to our software in the future. Data production is a batch process distributed across a large cluster of processors. We normally produce a new iteration of the complete database once per year, adding new samples to the collection and modifying variables as warranted. Because variables are harmonized across samples, any change in coding structure potentially affects all samples, necessitating re-processing the entire data collection.

A variety of checks aim to ensure the quality of the output data. Certain types of errors are logged by the data conversion program as it runs: values not accounted for in translation tables, values created via programming that lack labels, or assigned values wider than the designated width of the output field. As the data are run, the program records the frequencies of every variable in each sample. This information is used in the web interface as well as diagnostically. We collate the frequencies for each harmonized variable and compare the distributions across countries as well as focusing closely on change over time within countries. Because most processes are driven by metadata under the control of research staff, iterations to resolve data issues usually do not require hand-offs back to the programmers, which is

a significant productivity advantage of our approach. The constructed variables created via programming draw special attention, especially the family interrelationship variables identifying spouses and parents across records within households (Sobek and Kennedy 2009). That programming is complex and often requires custom solutions because of idiosyncrasies in the source data.

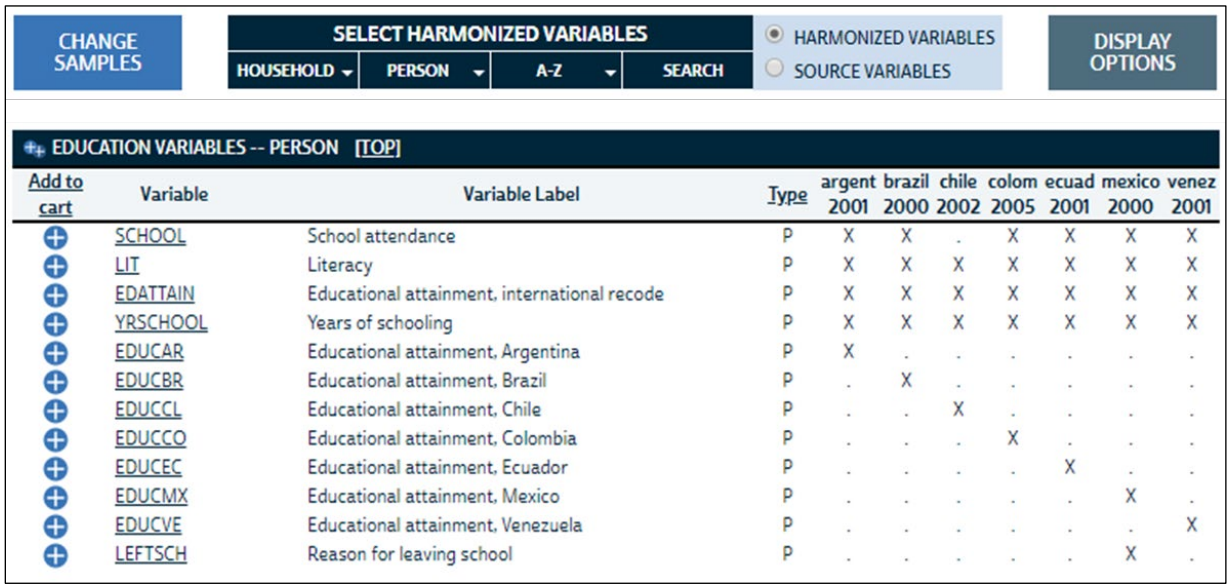
If we discover that data for a variable are clearly erroneous, we suppress it from appearing in the dissemination system via a metadata switch. We can suppress entire variables or a single sample within a harmonized variable. We do not systematically look for instances where data in two variables for a sample contradict one another, although such inconsistencies are sometimes revealed during the specification of variable universes. In general, IPUMS does not edit the data, aside from consolidating obviously erroneous values, like impossible ages or hours of work, into a single omnibus missing-value category. Where a sample has a large proportion of missing values in a variable, we note that in the harmonized description. The IPUMS data conversion program supports missing data allocation, and we have implemented that approach for other data collections, but we have not taken that step with the international census samples. As the sole distributor of many of these files, we tend to be conservative.

Dissemination

A multi-featured dissemination system is an essential component of our harmonization approach. Harmonized data are more complex than data from a single survey or census and they pose unique dissemination challenges. A key task for any user is data discovery: what samples are available and what variables do they contain. A one-dimensional list of variables becomes a two-dimensional grid in an integrated data collection. IPUMS has over 1700 harmonized variables and 40,000 unharmonized source variables. Most samples contain no more than one to two hundred variables, and there are hundreds of samples. This amounts to a large and relatively sparse matrix of variable availability, though some basic variables are almost universal. The IPUMS user interface offers several tools to help users explore the contents of the database.

Data exploration centers on variables. Variable browsing is segregated into harmonized and (unharmonized) source variable modes. A drop-down menu grouping variables into topics is the default mechanism for examining the database's contents. Alternatively, a Boolean variable search feature lets users specify the metadata components they wish to scan: descriptions, questionnaire text, variable labels, and value labels. Sample selection is another powerful data discovery tool. Most users have some geographic, and possibly temporal, scope in mind for their research. At any point while browsing, users can select samples of interest, winnowing the list of variables to only those that appear in at least one of the chosen datasets. Figure 3 shows the availability grid for the Education variable group after selecting a set of South American samples from the 2000 census round. Each row in the Figure is a variable, and the columns on the right represent the samples (Argentina 2001, Brazil 2000, etc.). An "X" indicates the variable is available in that sample. In this example, Chile 2002 is the only sample that lacks the first variable, school attendance. Every sample contains the harmonized educational attainment variable (EDATTAIN), but numerous compromises were involved in creating an internationally comparable classification on this topic. Consequently, each sample also has its own country-specific education variable that remains true to the national schooling system and its nomenclature.

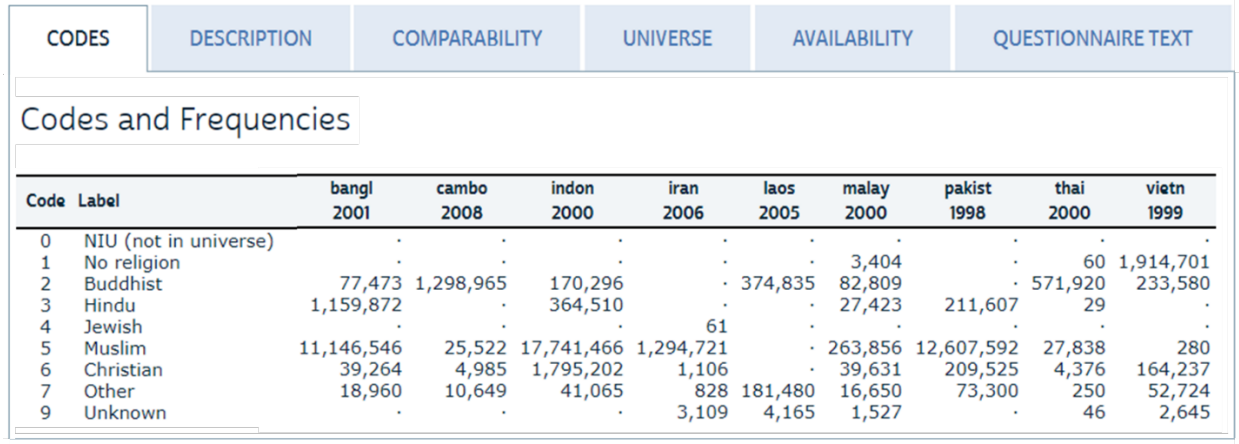
Figure 3. Variable browsing page: Education group



Add to cart	Variable	Variable Label	Type	argent 2001	brazil 2000	chile 2002	colom 2005	ecuad 2001	mexico 2000	venez 2001
⊕	SCHOOL	School attendance	P	X	X	.	X	X	X	X
⊕	LIT	Literacy	P	X	X	X	X	X	X	X
⊕	EDATTAIN	Educational attainment, international recode	P	X	X	X	X	X	X	X
⊕	YRSCHOOL	Years of schooling	P	X	X	X	X	X	X	X
⊕	EDUCAR	Educational attainment, Argentina	P	X	.	.	.	.	.	.
⊕	EDUCBR	Educational attainment, Brazil	P	.	X	.	.	.	.	.
⊕	EDUCCL	Educational attainment, Chile	P	.	.	X	.	.	.	.
⊕	EDUCCO	Educational attainment, Colombia	P	.	.	.	X	.	.	.
⊕	EDUCEC	Educational attainment, Ecuador	P	.	.	.	.	X	.	.
⊕	EDUCMX	Educational attainment, Mexico	P	.	.	.	.	.	X	.
⊕	EDUCVE	Educational attainment, Venezuela	P	.	.	.	.	.	.	X
⊕	LEFTSCH	Reason for leaving school	P	.	.	.	.	.	X	.

Clicking on a variable brings up its integrated documentation. Figure 4 displays the documentation page for Religion for a set of Asian countries for the 2000 census round. In this view, the codes and frequencies are displayed (the "Codes" tab for Religion). Note that sample filtering extends into the variable documentation; thus, only frequencies for the selected samples are displayed. For data exploration, category availability can be a critical consideration, allowing researchers to determine whether particular comparisons can be sustained without first downloading and analyzing the data. For example, none of these censuses includes categories for every major religion, with Laos only identifying Buddhism. The unweighted case counts offer additional guidance for users. Hindus are identified in five samples, but Thailand has too few cases for analysis.

Figure 4. Integrated variable documentation: Religion



Code	Label	bangl 2001	cambo 2008	indon 2000	iran 2006	laos 2005	malay 2000	pakist 1998	thai 2000	vietn 1999
0	NIU (not in universe)	·	·	·	·	·	·	·	·	·
1	No religion	·	·	·	·	·	3,404	·	60	1,914,701
2	Buddhist	77,473	1,298,965	170,296	·	374,835	82,809	·	571,920	233,580
3	Hindu	1,159,872	·	364,510	·	·	27,423	211,607	29	·
4	Jewish	·	·	·	61	·	·	·	·	·
5	Muslim	11,146,546	25,522	17,741,466	1,294,721	·	263,856	12,607,592	27,838	280
6	Christian	39,264	4,985	1,795,202	1,106	·	39,631	209,525	4,376	164,237
7	Other	18,960	10,649	41,065	828	181,480	16,650	73,300	250	52,724
9	Unknown	·	·	·	3,109	4,165	1,527	·	46	2,645

All related variable metadata is accessible through tabs from this unified variable screen, including the comparability discussion, universe information, and questionnaire text. The questionnaire text is in English, but the tab includes links to images of the original forms. As with the codes page, each of these elements is filtered based on sample selection. Without this filtering mechanism, users could easily become overwhelmed, negating the value of our voluminous documentation and increasing the likelihood of misinterpretations.

As users browse variables, they can add them to their data cart. When they are finished selecting variables and samples, they enter the cart and submit their data extract request. The extract can contain both harmonized variables that apply globally and sample-specific source variables. The IPUMS extract engine will build a single pooled dataset from the request, potentially including variables from hundreds of samples. Depending on the size of the request, the job can take a few minutes to a few hours. An automated email to the user indicates the extract is ready for downloading. Data can be produced as ASCII files with SAS, SPSS, and STATA syntax files. The data can also be produced directly as system files in the native format of the statistical packages. An R package and CSV output are also available.

Several options in the extract process are aimed at helping users manage the logistics of these potentially very large data extracts. Users can subset the data to include only certain cases, such as

persons aged 60 and older. More usefully for most research purposes, users can refine the subsetting feature to include *all residents in households* that contain any person aged 60 and older. The system will also draw a systematic subsample of each census at whatever density is requested, while adjusting the sample weights. A final option capitalizes on the hierarchical structure of the data and a set of constructed variables that identify the record numbers of each person's co-resident spouse, mother, and father. The system will attach the characteristics of a parent or spouse as a new variable on each person's record: for example, the employment status of each person's mother, or the educational attainment of their spouse.

IPUMS also provides an online data analysis system for users who lack software to analyze the microdata on their desktop, or who only require some quick tabulations. The online analysis software was developed by researchers at Berkeley and can produce a tabulation on most data files in a few seconds. It can also perform regression analysis and calculate confidence intervals and other statistical measures. Because the data are harmonized across countries, we are able to offer online tabulations that include multiple censuses, supporting even continent-wide comparisons in a single tabulation. We are not aware of any other online tabulation system that pools microdata in this way.

Conclusion

IPUMS is designed to address both the logistical and informational challenges posed by comparative analysis of largescale microdata. Data harmonization aims to resolve basic variable comparability issues, pairing like with like and applying consistent codes and labels. But some conceptual and substantive differences often persist. At a certain point, the researcher is best equipped to decide how particular census differences might affect their analysis and what should be done about them. But to make those determinations, researchers need information. We have tried to design a web system that provides the tools researchers require to make informed decisions while freeing them from the more mechanical tasks of managing data and collating documentation.

IPUMS is a metadata project as much as a data project. IPUMS was intended from the outset to be permanent research infrastructure (Ruggles, Hacker, and Sobek 1995; Ruggles, Sobek, and Gardner 1996). The data, documentation, and dissemination systems are all driven by the same metadata, which ensures that they always remain synchronized. In addition to the metadata types described above, a set of control files govern processing and display of variables, samples, and countries. Driving processes with metadata makes the project sustainable in the face of evolving technology. New web or data conversion software can be substituted as needed, while the labor-intensive metadata persists. Our primary tool for variable harmonization—the translation table—offers flexibility. We can never know the full universe of codes requiring harmonization, because new censuses are being conducted continuously; thus, variables will inevitably need periodic modification. The translation tables are essentially self-documenting and easy to amend to accommodate new categories.

The primacy of metadata in IPUMS allows the content specialists to control virtually all aspects of the data and documentation with which users interact. For metadata work, we use the simplest tools possible: spreadsheets and text editors. The software validates and pulls this metadata into a database, but for most tasks the research staff use Excel as their front-end, which provides access to many well-developed features that are useful during development, such as sorting, filtering, hiding columns, and commenting. Using familiar and relatively simple software whenever possible reflects our general strategy of minimizing the technical burden on the decision-makers while leaving the heavy lifting to the software developers.

The IPUMS web dissemination system is essential to our harmonization approach. Data discovery and access require specialized software to make the database useable. IPUMS is otherwise too large and heterogeneous to understand and manage. We write variable documentation with the features of the user interface in mind, and we continually seek ways to improve the system for researchers. The creation of source variables as a standardizing stage prior to harmonization was one such improvement. It made

the task of harmonization much more manageable for us while offering a means to provide users access to the full detail of the original datasets. Researchers can even use the source variables to deconstruct our harmonized classifications.

IPUMS is not limited to census data. We have applied the methods described here to a range of U.S. and global survey data collections as well (Sobek et al 2011; Ruggles 2014). Each of those databases also involve ex poste harmonization, but surveys present some different challenges as well as opportunities. On one hand, the survey collections have more variables and are usually conducted more frequently than censuses, producing logistical issues for metadata creation and data discovery. On the other hand, there are typically commonalities across the surveys within a collection due to "model" or infrequently modified questionnaires. Consequently, the surveys we have processed have generally required less intense harmonization than censuses and have not needed the intermediate source variable standardization stage. We have also been able to leverage consistency in variable names, labels, and codes to auto-populate translation tables in many cases. In the future, we hope to develop web tools for cross-collection data discovery, and perhaps someday develop a new processing stage to harmonize the harmonized data collections to each other.

REFERENCES

Esteve, A. and Sobek, M. 2003. Challenges and methods of international census harmonization. Historical Methods, 36: 66-79.

Cleveland, L., Davern, M. and Ruggles, S. 2011. Drawing statistical inferences from international census data Minnesota Population Center Working Paper, 2011-1.

Kugler, T.A., Manson, S.M. and Donato, J.R., 2017. Spatiotemporal aggregation for temporally extensive international microdata. Computers, Environment and Urban Systems, 63: 26-37.

McCaa, R., Ruggles, S. and Sobek, M. 2011. IPUMS-International statistical disclosure controls. In J. Domingo-Ferrer and E. Magkos, Editors, Privacy in Statistical Data 2010, LNCS 6344. Berlin and Heidelberg: Springer Verlag. p. 74-84.

Meier, A., R. McCaa, and D. Lam. 2011. Creating statistically literate global citizens: The use of IPUMS-International integrated census microdata in teaching. Statistical Journal of the International Association for Official Statistics 27: 145-156.

Minnesota Population Center. Integrated Public Use Microdata Series, International: Version 7.2 [dataset]. Minneapolis, MN: IPUMS, 2019. https://doi.org/10.18128/D020.V7.2

Ruggles, S. 2014. Big microdata for population research. Demography, 51: 287-297.

Ruggles, S., Hacker, J.D. and Sobek, M. 1995. Order out of chaos: The Integrated Public Use Microdata Series. Historical Methods, 28:33-39.

Ruggles, S., McCaa, R., Sobek, M. and Cleveland, L. 2015. The IPUMS Collaboration: Integrating and Disseminating the World's Population Microdata. Journal of Demographic Economics, 81: 203-216.

Ruggles, S., Roberts, E., Sarkar, S. and Sobek, M. 2011. "The North Atlantic Population Project: Progress and Prospects." Historical Methods, 44: 1-6.

Ruggles, S., Sobek, M. and Gardner, T. 1996. Disseminating historical census data on the World Wide Web. IASSIST Quarterly, 20:4-18. http://www.iassistdata.org/downloads/iqvol203ruggles.pdf

Sobek, M. and Cleveland, L. 2017. IPUMS approach to harmonizing international census and survey data. United Nations Economic Commission for Europe, Conference of European Statisticians, Working Paper 31.

Sobek, M. and Kennedy, S. 2009. The development of family interrelationship variables for international census data. Minnesota Population Center Working Paper, 2009-2.

Sobek, M., Cleveland, L., Flood, S., Ruggles, S. and Schroeder, M. 2011. Big data: Large-scale historical infrastructure from the Minnesota Population Center. Historical Methods, 44: 61-68.

Sobek, M., Hindman, M., and Ruggles, S. 2007. Using cyber-resources to build databases for social science research. Minnesota Population Center Working Paper Series, 2007. #07-01.

United Nations. 2017. Principles and recommendations for population and housing censuses, Revision 3. Department of Economic and Social Affairs, Statistics Division.