IPUMS approach to harmonizing census and survey microdata

Submitted for consideration for the UAPS Conference 2019

Lara Cleveland and Matthew Sobek
IPUMS Center for Data Integration
University of Minnesota
clevelan@umn.edu

Introduction

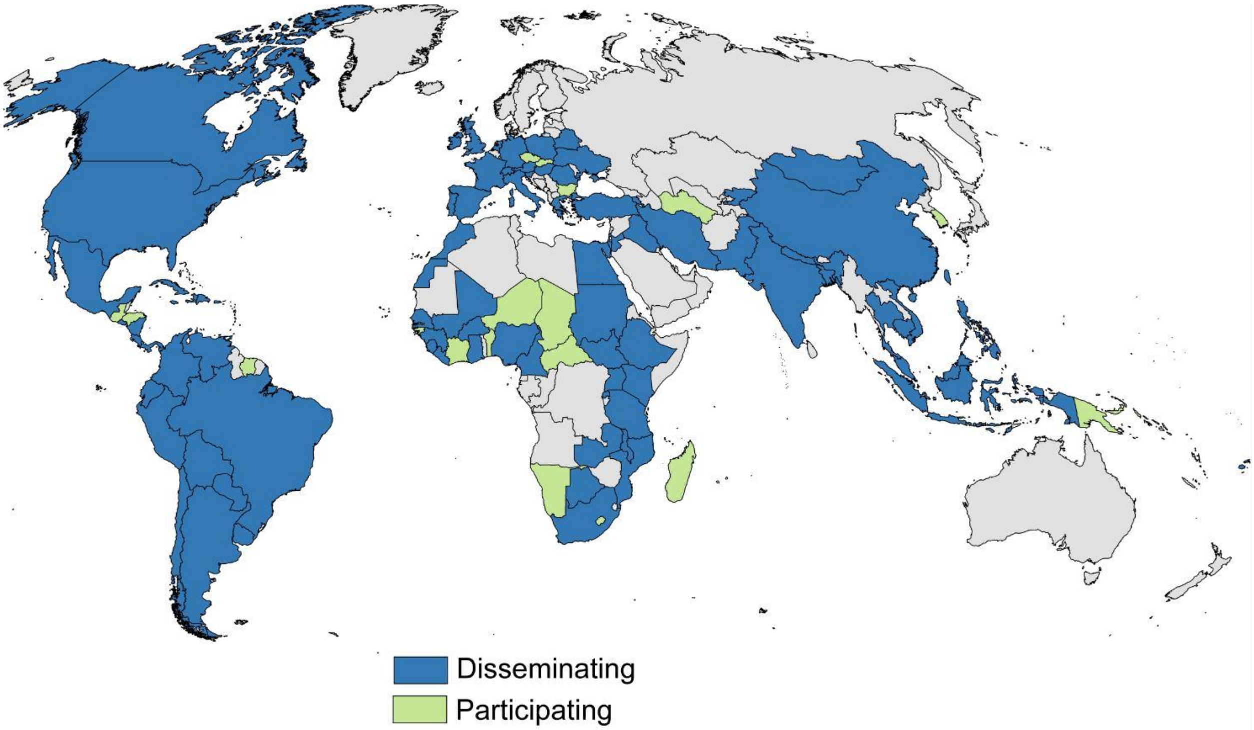
IPUMS is the world's largest collection of population microdata available for research and education. The project integrates census data from 85 countries into one consistent database. The signature feature of IPUMS is to harmonize variables across countries and over a fifty year period, so the same code has the same meaning in all times and places. The aim is to facilitate comparative research by reducing the cognitive and logistical burden on researchers, enabling them to focus on analysis.

The anonymized microdata files provided to IPUMS by National Statistical Offices arrive coded into a wide variety of classification schemes dating from when they were originally processed in the different countries. Most variables simply report the categories that were listed as response options on the particular census questionnaire from which they were derived. There is no standardization across countries and little consistency within countries over time. Some countries adhere to international classifications when such standards exist, while others modify or ignore them. To enable comparative research requires coping with this empirical reality of pre-classified microdata.

About IPUMS

IPUMS is composed of census microdata: each record is a person, and all of their individual characteristics are known. Microdata allow researchers to create tabulations never envisioned by the collectors of the data, and they enable sophisticated multivariate modelling. IPUMS currently includes data for 672 million individuals recorded in over 300 censuses taken since 1960 (Ruggles et al 2015). Most countries provide multiple censuses, enabling study of change over time both nationally and internationally. IPUMS is the sole public dissemination mechanism for many, if not most, of the countries in the database. To ensure confidentiality, the data do not include names or low-level geographic identifiers. As a further reinforcement of privacy, the data are samples, typically comprising one to ten percent of the national population. Prospective users must apply for access, and over 14,000 researchers have been registered. Figure 1 shows the geographic coverage of IPUMS, distinguishing between countries for which data are currently being disseminated and others that have yet to be processed.

Figure 1. IPUMS Geographic Coverage



A web dissemination system allows users to browse the contents of the database and construct custom data extracts that pool data from multiple countries and time periods into a single file. The user downloads the file -- typically containing some millions of records and twenty to thirty variables -- to their desktop for analysis. Through the web system, researchers have access to detailed documentation for each variable; including comparability discussions, codes, frequencies, and other information. The website is accessible at https://international.ipums.org and through the broader portal www.ipums.org, which serves the full suite of harmonization projects developed by IPUMS for various data collections (Sobek 2011).

A critical characteristic of microdata lies in the categorical detail it retains at the individual level. It is this detail that makes it feasible to harmonize the data across countries and over time. The tabulated data that are the traditional product of each census often cannot be meaningfully harmonized cross-nationally, because of decisions built into their construction. With IPUMS, researchers can devise custom tabulations using the full detail of the microdata while imposing consistent population universes across samples. Microdata will also support the kinds of multivariate analyses conducted by most academic and policy researchers. The data are cross-sections in time; it is not possible to link people across censuses.

The IPUMS samples incorporate most of the detail from the original census questionnaires. All censuses have basic demographic information such as age, sex, and marital status. Nearly as universal are socioeconomic variables, such as education, employment status, and occupation. There is considerable topical variation beyond these, but questions on migration, ethnicity, disability, and fertility are also broadly asked (Sobek 2016). Most censuses, particularly in the developing world, have information about the dwelling as well, such as construction materials, plumbing, utilities, and household assets. Figure 2 provides a high-level summary of the topical coverage commonly available in censuses.

Figure 2. Common Census Topics

Person	Dwelling
Basic demographics	Location
Marriage	Home ownership
Family structure	Construction materials
Fertility and mortality	Sanitation
Ethnicity and language	Electricity and utilities
Education	Rooms and bedrooms
Work	Household assets
Disability	
Migration	

Harmonization overview

IPUMS harmonizes variables across the entire database. There are three elements to variable harmonization: applying consistent codes across samples, determining labels for those codes, and collating integrated variable descriptions that speak to issues not sufficiently conveyed by codes and labels.

The central harmonization challenge is to equate codes that have the same meaning for a variable that is common across samples. This is fundamentally a metadata issue. One must understand the meaning of the codes, which is conveyed by their labels, the coding structures, and by the deeper context of the census questionnaire text and enumerator instructions. Each of those elements poses challenges. The labels provided with census files are often shorthand for more complex concepts or combinations of items. They may have been created ad hoc during processing, and in most cases they have been translated out of their original language into English at some cost to their precise meaning. Coding schemes often have structure, where the meaning of a particular category can only be understood in the broader context of the classification. This is especially true for residual categories, such as "Other relatives", whose meaning is defined by the other categories that are enumerated in the classification. Finally, much meaning is embedded in how the census question was worded and in the instructions given to the census enumerators regarding the question. For example, some countries restrict the status of being "married" to only legal marriages, while others make allowance for "common law," custom, or other variations. Those distinctions are often not reflected in the value labels, and may only be discoverable from the questionnaire or instructions.

Population data harmonization ultimately depends on informed human judgement. Computers can help greatly with the logistics, but they can provide only limited leverage equating the meanings of international census data, which depend so much on context. IPUMS has nevertheless written a great deal of software to assist with the harmonization process. In most cases, researchers manipulate metadata to standardize and harmonize the data, with the software being driven by the metadata. A description of that process follows below.

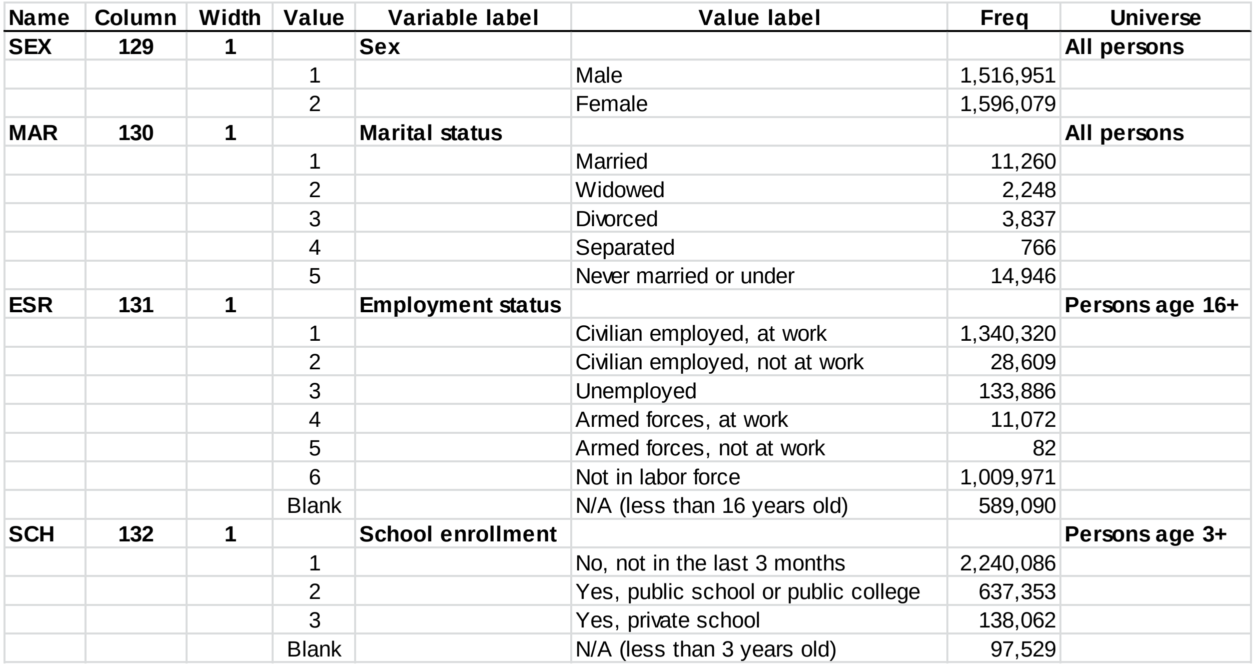
The census data provided to IPUMS come in many formats with varying documentation in many languages. The categorical variables in recent censuses often reflect the influence of international standards and recommendations, but countries may choose to modify or ignore them. The older data are less regularized in every respect. Harmonizing data from such disparate source material is a complicated process, and we break it down into a series of discrete steps to make it manageable and efficient. To the extent possible, we strive for an industrial as opposed to a craft model of production.

Data standardization

Before data processing can commence, one must understand the data structure. We require basic metadata to interpret the files: the relationship between data records, the linking keys, names and locations of variables, and labels for categorical variables. These metadata must be translated into English, as necessary, before we begin. We cannot retain all possible language skills on our team, nor do we want particular staff to have exclusive responsibility for particular samples.

IPUMS metadata development begins with the creation of a data dictionary for each dataset. An IPUMS data dictionary is much like a codebook, but it contains more information and in a more structured format suited to machine processing. IPUMS software is designed to read this metadata structure. Figure 3 shows a small part of a data dictionary. It records each source variable's name, location in the data file, labels for variables and values, frequencies for each value, universe of respondents, and any other fields needed to fully document the data or control data processing, such as indicating string fields or implied decimal places. Some of these fields may not be immediately known, but are added later during processing and examination.

Figure 3. Data dictionary



Name	Column	Width	Value	Variable label	Value label	Freq	Universe
SEX	**129**	**1**		**Sex**			**All persons**
			1		Male	1,516,951	
			2		Female	1,596,079	
MAR	**130**	**1**		**Marital status**			**All persons**
			1		Married	11,260	
			2		Widowed	2,248	
			3		Divorced	3,837	
			4		Separated	766	
			5		Never married or under	14,946	
ESR	**131**	**1**		**Employment status**			**Persons age 16+**
			1		Civilian employed, at work	1,340,320	
			2		Civilian employed, not at work	28,609	
			3		Unemployed	133,886	
			4		Armed forces, at work	11,072	
			5		Armed forces, not at work	82	
			6		Not in labor force	1,009,971	
			Blank		N/A (less than 16 years old)	589,090	
SCH	**132**	**1**		**School enrollment**			**Persons age 3+**
			1		No, not in the last 3 months	2,240,086	
			2		Yes, public school or public college	637,353	
			3		Yes, private school	138,062	
			Blank		N/A (less than 3 years old)	97,529	

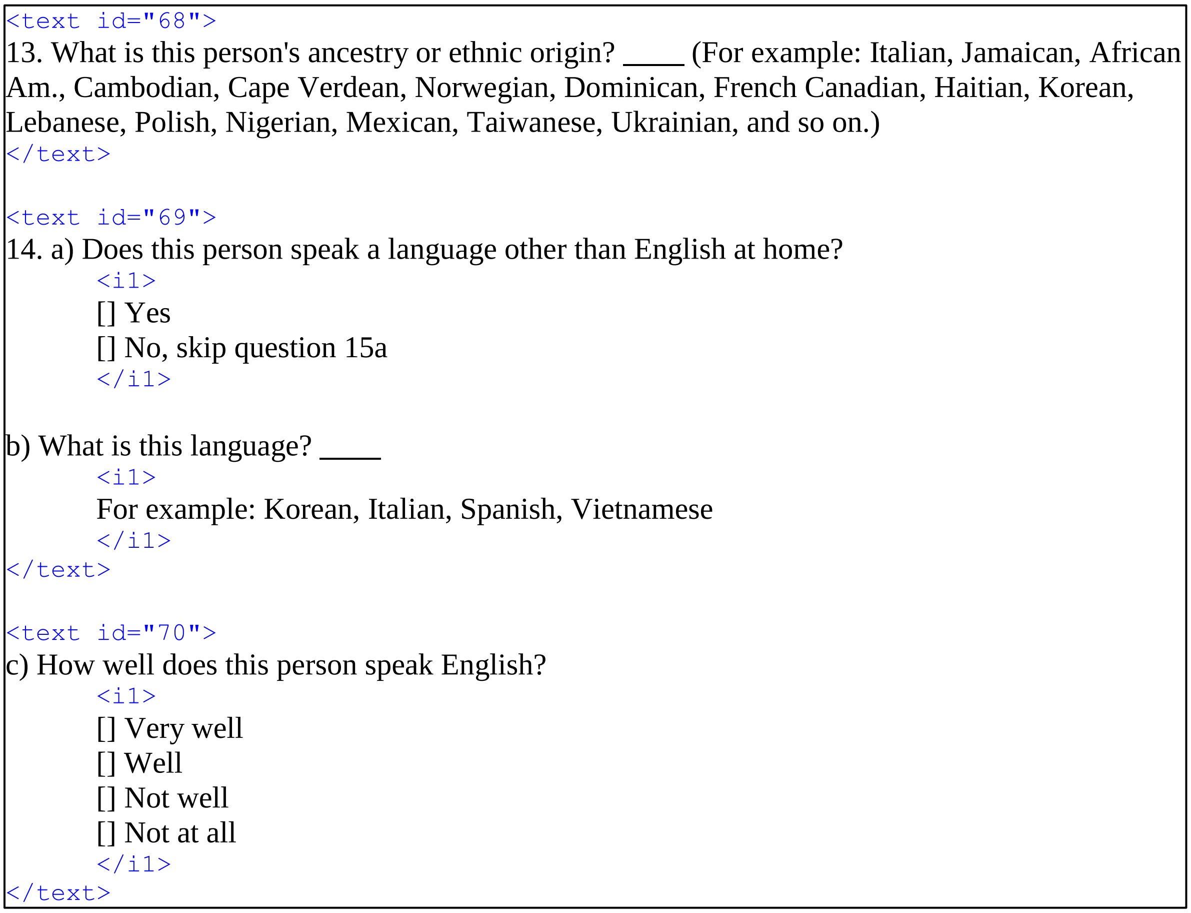
The first stage of processing is to convert the source datasets into a common format. We turn all datasets into fixed-format ASCII files with a hierarchical structure: each household record is followed by multiple person records representing its members. We receive data in many formats that might require merging separate household and person files, converting out of native SPSS or Stata format, reorganizing files with complex geographic hierarchies, or other manipulations. In the process of regularizing the data structures we create some common technical variables useful for our system.

Custom programming is often required at this stage, because unique situations commonly arise and errors may be uncovered. As we modify the data, any changes to variables or record layout are recorded in the data dictionary, which evolves to stay in sync with the data file. Once formatting is completed, the data is in a form understood by the rest of our data transformation, diagnostic, and web software.

At this point we have processable input data files, but we are not yet ready to harmonize the variables. Early in the history of IPUMS, we discovered that harmonizing variables from dozens of international organizations directly from the reformatted input data was too difficult. We therefore inserted additional steps to standardize the variables first. The goal is clean, well-documented source variables to use as input for harmonization. Every input variable is analyzed to confirm the universe of respondents, which is recorded in the data dictionary. We also perform limited data recoding to regularize the variables. For example, we combine stray values that are clearly data errors into single missing categories, or we separate meaningful zeroes from non-responses where they can be logically sorted out. The aim at this point is minimalism. Data are only recoded to ensure fully documented, clean variables as input to the harmonization stage. No meaningful information is lost. This low-level variable recoding is governed by additional entries in the data dictionary, with supplemental programming as needed.

The final part of variable standardization involves connecting the source variables via metadata with their associated text in the census questionnaire and instructions. This information is necessary to fully understand the variable and is crucial during harmonization. The task is to convert pdfs and other static documentation into usable, machine-actionable metadata. To this end, all census questionnaires and instructions are translated into English and converted into a custom XML format. Figure 4 shows a part of one such marked-up questionnaire. Within the XML, each question and block of text is assigned an index number ("text id"). IPUMS researchers insert the relevant index numbers into the data dictionary for each variable to associate it with the questionnaire language that produced it. Having systematized this material, it can be compiled on demand using software, for both internal use and in the web dissemination system.

Figure 4. XML-tagged questionnaire



```
<text id="68">
13. What is this person's ancestry or ethnic origin? ____ (For example: Italian, Jamaican, African
Am., Cambodian, Cape Verdean, Norwegian, Dominican, French Canadian, Haitian, Korean,
Lebanese, Polish, Nigerian, Mexican, Taiwanese, Ukrainian, and so on.)
</text>

<text id="69">
14. a) Does this person speak a language other than English at home?
        <i1>
        [] Yes
        [] No, skip question 15a
        </i1>

b) What is this language? ____
        <i1>
        For example: Korean, Italian, Spanish, Vietnamese
        </i1>
</text>

<text id="70">
c) How well does this person speak English?
        <i1>
        [] Very well
        [] Well
        [] Not well
        [] Not at all
        </i1>
</text>
```

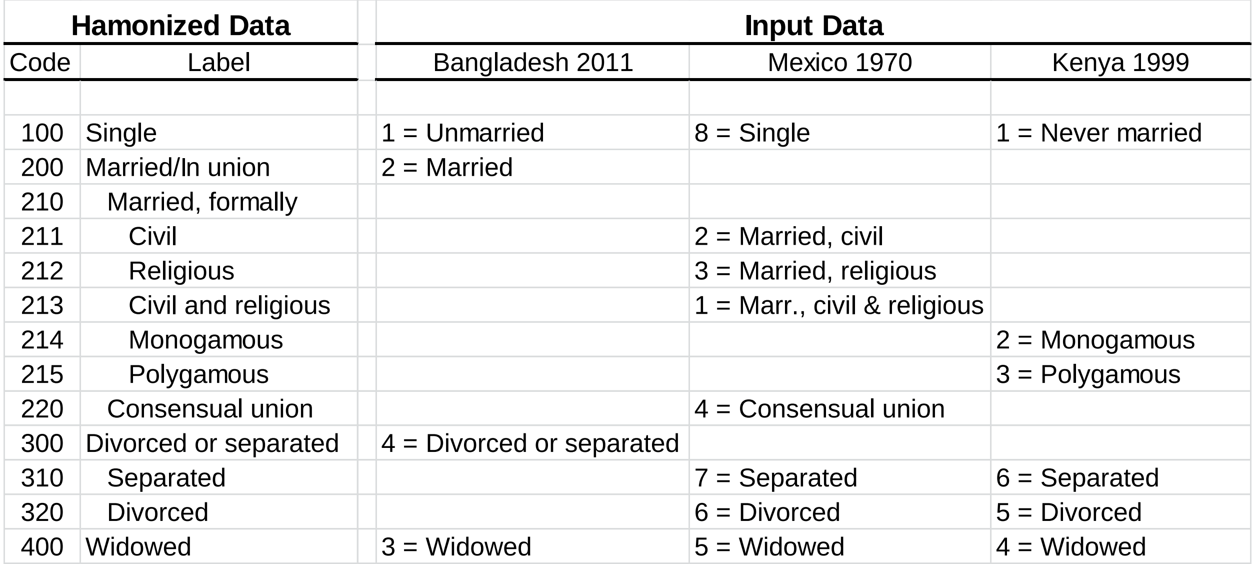
Only after the input variables have been standardized and fully documented are we ready to proceed to harmonization. To the extent possible, the previous stages are as rules-driven and mechanistic as possible, trying to limit the scope of human decision making. No comparisons are made across datasets during standardization. Harmonization is a more creative activity requiring considerable judgement and problem-solving.

Variable harmonization

At the highest level, harmonization requires determining which variables are conceptually the same across datasets (Esteve and Sobek 2003). Beyond variable names and labels, such determinations may require referring to codes, value labels, text of census questions, category frequencies, or other metadata. This is sometimes a judgement call for the harmonizer, who must ask whether combining variables with differing shades of meaning is likely to mislead researchers trying to interpret the data. Even if the concepts appear equivalent, an additional issue concerns the fundamental compatibility of the classifications. For example, continuous variables may be coded into incompatible value ranges, or different censuses may group response items in overlapping ways that defy harmonization.

The signature activity of data integration is to harmonize variable codes and labels across data samples. Our primary device for achieving this is a "translation table" like the one for Marital Status depicted in Figure 5. The leftmost columns contain the harmonized output values and their labels. Each column on the right side documents every value that exists in one of the input datasets being harmonized: in this case census samples from three developing countries: Bangladesh, Mexico, and Kenya. 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. The work is performed by a researcher using the tools we have developed specifically for this process. In broad strokes, the process is as follows: a researcher identifies the source variables, a program directly inserts the input values into the translation table from the appropriate data dictionaries, and a researcher then aligns the codes and assigns output codes and labels (the "harmonized data" columns on the left). Thus, the original codes "1: Unmarried", "8: Single", and "1: Never married" are all aligned and will be recoded to the internationally harmonized IPUMS output code "100: Single". This sort of semantic integration is intellectual labor that no computer program can perform. It 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.

Figure 5. Translation Table (Marital Status)



Hamonized Data		Input Data		
Code	Label	Bangladesh 2011	Mexico 1970	Kenya 1999
100	Single	1 = Unmarried	8 = Single	1 = Never married
200	Married/In union	2 = Married		
210	Married, formally			
211	Civil		2 = Married, civil	
212	Religious		3 = Married, religious	
213	Civil and religious		1 = Marr., civil & religious	
214	Monogamous			2 = Monogamous
215	Polygamous			3 = Polygamous
220	Consensual union		4 = Consensual union	
300	Divorced or separated	4 = Divorced or separated		
310	Separated		7 = Separated	6 = Separated
320	Divorced		6 = Divorced	5 = Divorced
400	Widowed	3 = Widowed	5 = Widowed	4 = Widowed

Our harmonization of variables is designed to meet two goals: 1) retain all the detail provided in the original samples; and 2) provide a truly 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 marital status in Figure 5 illustrates the composite coding approach. In this example, the first digit of marital status has four categories consistently available in all samples: 1) single, 2) married/in union, 3) divorced or separated, and 4) widowed. The distinction between divorced and separated is not maintained in all samples, so these categories are combined at the fully comparable first digit. At the second digit, we distinguish divorced and separated persons in the samples with that information, as well as formal marriages and consensual unions. The third and final digit differentiates among types of marriages (civil, religious, polygamous) available for select countries only. 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 and offers the assurance that most comparability issues are resolved.

We also leverage the questionnaire tagging (described above) to inform the integration process. As soon as the source variables for a harmonized variable are identified, our software is able to compile the questionnaire text from all the samples. Thus, the research staff is able to make their variable and code harmonization determinations with the most critical information readily at hand, combined in one view. If the actual question wording for a variable indicates a significant conceptual difference between samples, we create a separate variable to minimize the likelihood of user error.

Our approach to variable harmonization demonstrates an underlying principle in our integration methods. Our entire system represents what might be termed a metadata-centric approach, in which the research staff manipulates relatively simple but highly-structured documents that drive the data processing and web software. From these documents we generate a unique XML markup that identifies all elements necessary to guide the recoding and documentation of variables and to associate each variable with its relevant enumeration materials. The data, documentation, and dissemination systems are all driven by the same metadata, which ensures that they always remain synchronized.

The translation tables exemplify this metadata approach to data management and dissemination. We do not write recode statements in code, 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 future. The metadata-driven translation tables provide a practical solution to this challenge.

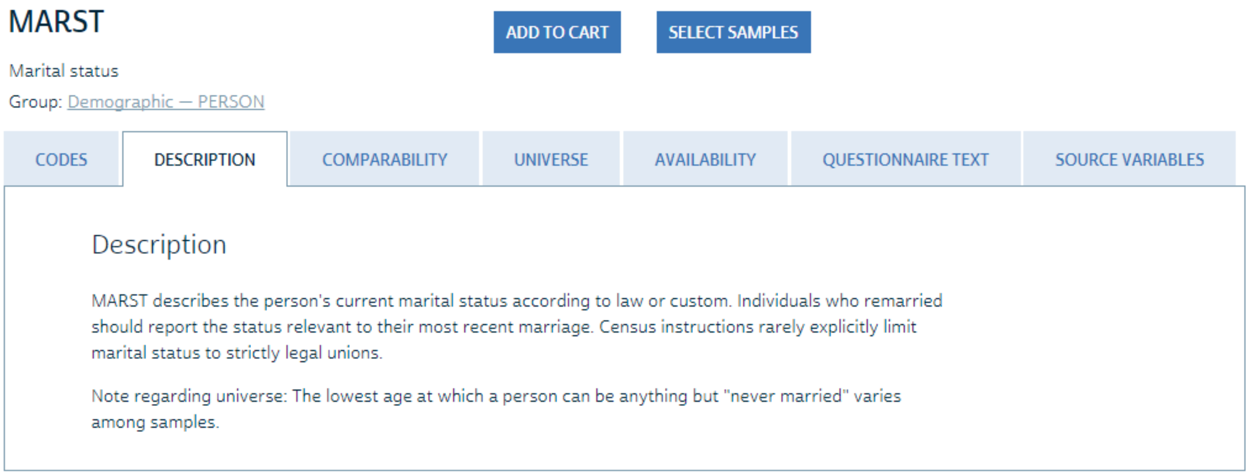
The custom IPUMS data conversion program reads the translation tables to produce the integrated output data. There are, of course, some instances where translation tables cannot accommodate the logic required to recode a variable, and variable-level programming is required; for example, for recoding continuous numeric variables like income into categories or combining multiple input variables. The data conversion program has the capacity to manipulate the data in any way required.

Harmonized documentation

Variable harmonization involves more than harmonizing codes. New documentation must be written for each integrated variable and made accessible to users. Because integrated variables have time and space dimensions, a key aspect of the documentation is to highlight any comparability issues that arise across samples. One area of focus is to indicate for users wherever changes in question wording may potentially cause subtle differences in meaning, even where the codes and labels look otherwise compatible. Changes in the universe of people who were asked the question are another common source of comparability issues. In these cases the primary aim of the description text is to direct the user's attention to the collated questionnaire text or universe statements for the variable. Our goal is to empower the researcher, who must ultimately decide if the issues that remain after harmonization are relevant to their analysis.

Because variable documentation is so critical to proper use of harmonized data, the IPUMS web dissemination system is an integral component of our approach. There is no avoiding the reality that harmonized data are simply more complex than discrete datasets. Users need better tools than pdf files and labels to understand and properly use the data. The IPUMS system lets them filter only the samples of interest and browse variables in an information-rich environment. There is, of course, no way to force researchers to avail themselves of the potential of the web system to inform their work, but we strive to make it as easy as possible. Figure 6 shows the Marital Status variable page in the web dissemination system. The series of tabs allow the user to explore all the metadata associated with the harmonized variable from one viewing pane.

Figure 6. Harmonized Variable Page in Web System: Marital Status



MARST

Marital status

Group: Demographic — PERSON

Description

MARST describes the person's current marital status according to law or custom. Individuals who remarried should report the status relevant to their most recent marriage. Census instructions rarely explicitly limit marital status to strictly legal unions.

Note regarding universe: The lowest age at which a person can be anything but "never married" varies among samples.

Harmonization challenges

The Marital Status example above exemplifies our approach to harmonization, but situations arise in the global census data that require alternative strategies. The remainder of this paper describes some of those scenarios and how we address them.

When we harmonize a variable we refer to any international standards that might exist for that topic. We particularly find useful the United Nations Principles and Recommendations for Population and Housing Censuses, which influences how many countries choose to ask certain questions (United Nations 2007). Unfortunately, many countries ignore this advice, and others appear to adopt it only loosely. But any standards are welcome. For our purposes, the UN principles also provide guidance about the salient features around which a harmonized variable might be organized. Some more complex census items, like occupation and industry, have well established international classifications used by a subset of countries every decade. Over the years, however, even those classifications have evolved, so there is never a time-invariant system for our purposes.

IPUMS greatly appreciates the application of standards in census questions and classification, but in the final analysis we must deal with the empirical reality of the data we are given. It is a truism that the least detailed classification among the input variables dictates the overall coding scheme of the harmonized variable. You can often recode more detailed variables to match simpler classifications, but one cannot add detail to variables that don't have it. In practice, this means the first digit of most harmonized variables is governed by the simplest classifications. But applying pure logic to harmonization can sometimes lead to variables that are hard to understand and use. Perhaps a sample(s) must be left out of the variable, or a category must be coded in a way that requires some caveat in the documentation.

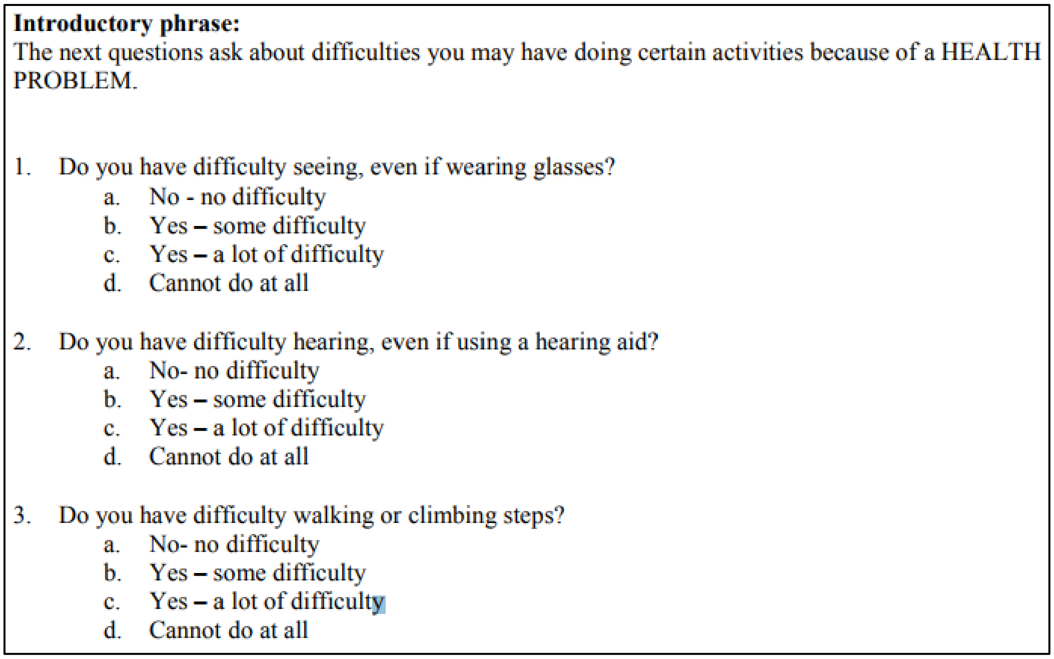
Literacy. Some variables require virtually no recoding to harmonize categories. Literacy and School Attendance are simple binary variables in nearly all countries, and there is little to do other than align the "no" and the "yes" responses. Despite their simplicity, however, there are definitional differences that cannot be conveyed via category labels. For example, some censuses define literacy as the ability to read and write a small paragraph, some use a threshold of years of schooling, and other censuses impose no objective standard. Whether such differences are important for a particular analysis is a question for the researcher. The only practical way to indicate such nuance is via the variable's comparability discussion and the feature to compile the questionnaire text.

Employment. Employment Status offers a similar challenge, but with a more concrete definitional difference. The variable is amenable to composite coding such as we employ for Marital Status: the first digit indicates employed, unemployed, and inactive persons; and trailing digits retain detailed categories whose availability varies across samples. But underlying the coding structure are differences in the reference period between countries. Different censuses assess employment status at the moment of the census, over a period of a week, or as an average over a longer time span. For some issues, like seasonal variations, the reference period can be important. The variable documentation must carry this information. The alternative is to create a set of parallel variables for the differing reference periods, but that would impose different costs on users.

Disability. Disability Status poses a more difficult challenge. Like literacy, disability is essentially a set of one or more binary variables (blindness, mobility impairment, etc.). But there are clearly cultural and census instruction differences at work within the data. Many censuses provide guidance to enumerators regarding what constitutes a disability, such as how to interpret loss of one eye or the need for a hearing aid; but other censuses provide little or no guidance. At some point, responses presumably depend on cultural norms. Even within countries where there are no discernable changes in question wording, the incidence of disability in a population can vary notably from one census to the next. In short, full comparability of disability statistics is difficult to attain under any circumstances.

To further complicate matters, there was a shift in the early 21st century toward adoption of the U.N. Washington Group set of questions on disability (Madans, Loeb, and Altman 2011). The new questions are intended to provide tightly comparable data across countries, aiming to identify functional limitations that produce social exclusion. Many countries in the 2010 census round have adopted the new question wording, shown in Figure 7, which employs the terminology of "some" or "a lot" of "difficulty" doing the particular activity. The creation of a world standard is laudable and should produce better and more comparable statistics among adopting countries. But equating degree of difficulty with older censuses is difficult, even within a single country. Statistical analysis suggests a disjuncture occurs when the new questions are imposed, which can yield much higher disability rates. In IPUMS, our approach has been to combine disability variables, interpreting "a lot" of difficulty as most comparable to the traditional questions. We include strong language in the variable comparability discussion warning researchers to be careful. However, as the Washington Group adherents have become more numerous, and as we've learned more about the issue, we now think we should create a distinct set of disability variables that adhere to the new approach. This would emphasize their difference from older samples and countries still using traditional questions, and it would highlight the high degree of comparability among the set of countries using the new standard.

Figure 7. Census Disability Questions Endorsed by the Washington Group (partial list)



Introductory phrase:
The next questions ask about difficulties you may have doing certain activities because of a HEALTH PROBLEM.

1. Do you have difficulty seeing, even if wearing glasses?
   a. No - no difficulty
   b. Yes – some difficulty
   c. Yes – a lot of difficulty
   d. Cannot do at all

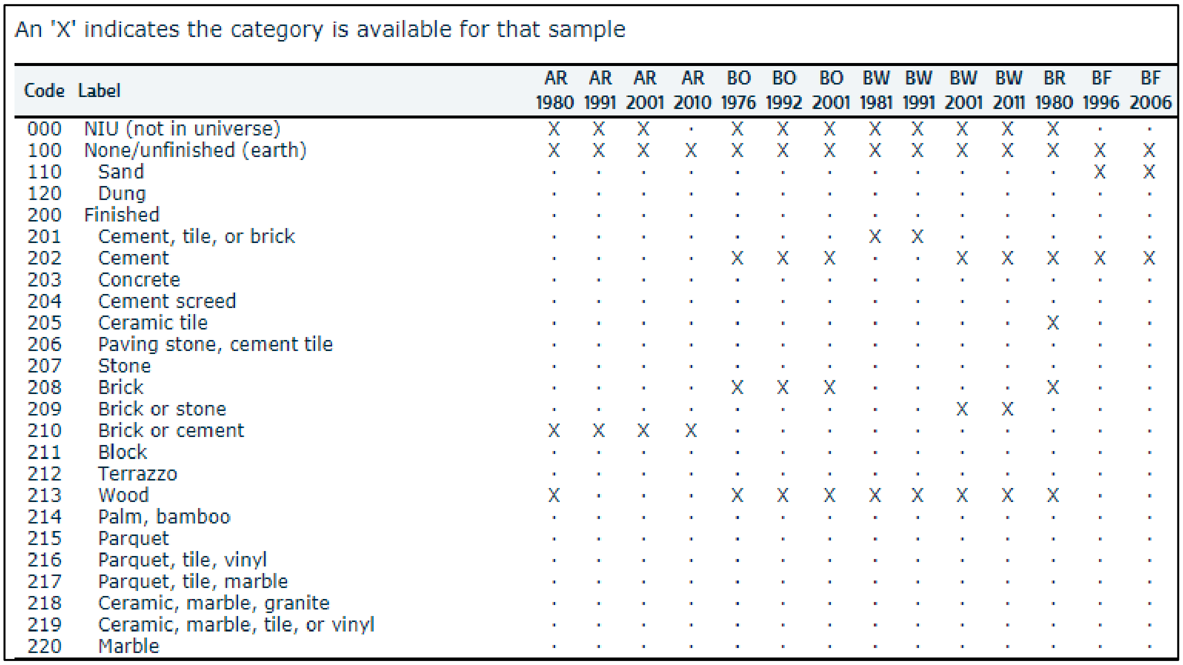
2. Do you have difficulty hearing, even if using a hearing aid?
   a. No- no difficulty
   b. Yes – some difficulty
   c. Yes – a lot of difficulty
   d. Cannot do at all

3. Do you have difficulty walking or climbing steps?
   a. No- no difficulty
   b. Yes – some difficulty
   c. Yes – a lot of difficulty
   d. Cannot do at all

Dwellings. Housing variables pose some of the more difficult harmonization challenges. Dwelling materials for floors, walls, and roofs can be highly localized, with terminology varying by language. Material types can be grouped together in ways that straddle groupings in other countries, defying prospects for strict logically nesting. For floors, a few major materials contain much of the variation: wood, concrete, stone, brick. Some terms, like "tile," are ambiguous. It is also not always clear where an unlisted material might be combined with others in the source data, given the limitations of the labels and number of categories available on the questionnaire.

Despite these issues, researchers would surely benefit from being able to manage a single variable with a lot of variation as opposed to many individual variables. The IPUMS approach in these cases might be termed partial harmonization. We concluded that the most useful distinction for most users -- and which is easily achievable in terms of consistent classification -- is to make the first digit distinguish only between unfinished (dirt) floors and finished floors. That binary distinction at the first digit captures key variation in terms of sanitation and socioeconomic status. The types of finished floors are grouped together as best as possible, but we do not use the second digit to suggest there is any structure to the 35 categories of finished floors. Thus, users have the data for all countries in one variable and access to all the original labels, with minimal modification. Figure 8 shows a snippet of the Floor variable codes page in the IPUMS dissemination system -- with each of the columns on the right representing a census, and the "X"s indicating the availability of the category for each sample. For analyses that require distinctions beyond finished-unfinished, the burden is on the user to group the codes as necessary. We take a similar approach with walls, roofs, and cooking fuel, roughly grouping categories and leaving the full original category labels in place. One might call this "nominal" harmonization, in that categories with the same label are assigned the same code, but their full unspoken contents may differ somewhat even within those categories.

Figure 8. Floor Variable Codes Page (partial)

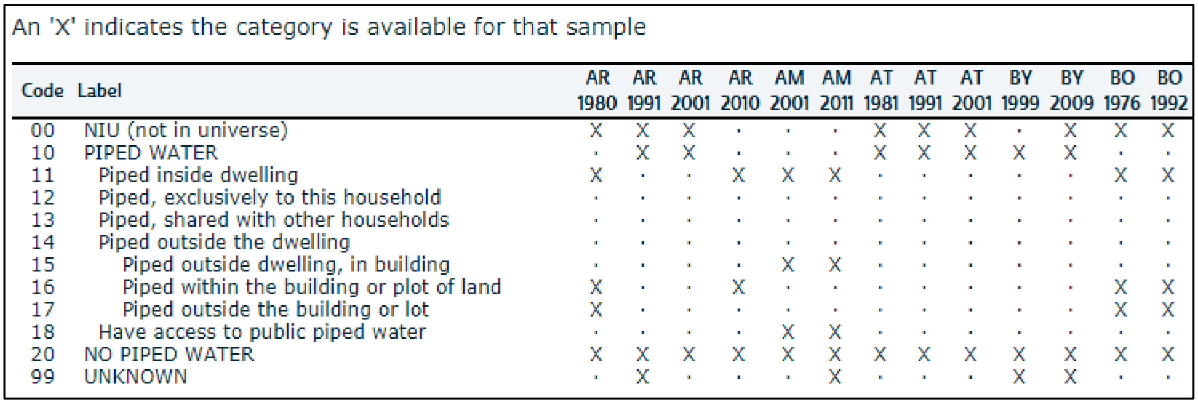


An 'X' indicates the category is available for that sample

Code	Label	AR	AR	AR	AR	BO	BO	BO	BW	BW	BW	BW	BR	BF	BF
		1980	1991	2001	2010	1976	1992	2001	1981	1991	2001	2011	1980	1996	2006
000	NIU (not in universe)	X	X	X	·	X	X	X	X	X	X	X	X	·	·
100	None/unfinished (earth)	X	X	X	X	X	X	X	X	X	X	X	X	X	X
110	Sand	·	·	·	·	·	·	·	·	·	·	·	·	X	X
120	Dung	·	·	·	·	·	·	·	·	·	·	·	·	·	·
200	Finished	·	·	·	·	·	·	·	·	·	·	·	·	·	·
201	Cement, tile, or brick	·	·	·	·	·	·	·	X	X	·	·	·	·	·
202	Cement	·	·	·	·	X	X	X	·	·	X	X	X	X	X
203	Concrete	·	·	·	·	·	·	·	·	·	·	·	·	·	·
204	Cement screed	·	·	·	·	·	·	·	·	·	·	·	·	·	·
205	Ceramic tile	·	·	·	·	·	·	·	·	·	·	·	X	·	·
206	Paving stone, cement tile	·	·	·	·	·	·	·	·	·	·	·	·	·	·
207	Stone	·	·	·	·	·	·	·	·	·	·	·	·	·	·
208	Brick	·	·	·	·	X	X	X	·	·	·	·	X	·	·
209	Brick or stone	·	·	·	·	·	·	·	·	·	X	X	·	·	·
210	Brick or cement	X	X	X	X	·	·	·	·	·	·	·	·	·	·
211	Block	·	·	·	·	·	·	·	·	·	·	·	·	·	·
212	Terrazzo	·	·	·	·	·	·	·	·	·	·	·	·	·	·
213	Wood	X	·	·	·	X	X	X	X	X	X	X	X	·	·
214	Palm, bamboo	·	·	·	·	·	·	·	·	·	·	·	·	·	·
215	Parquet	·	·	·	·	·	·	·	·	·	·	·	·	·	·
216	Parquet, tile, vinyl	·	·	·	·	·	·	·	·	·	·	·	·	·	·
217	Parquet, tile, marble	·	·	·	·	·	·	·	·	·	·	·	·	·	·
218	Ceramic, marble, granite	·	·	·	·	·	·	·	·	·	·	·	·	·	·
219	Ceramic, marble, tile, or vinyl	·	·	·	·	·	·	·	·	·	·	·	·	·	·
220	Marble	·	·	·	·	·	·	·	·	·	·	·	·	·	·

Dwelling water supply poses the challenge of dueling concepts among the source variables. The various censuses are oriented to a number of differing considerations: exclusive access to the water supply, piped water into the dwelling versus outside it, public piped water, and the ultimate source of the water (e.g., lake, river, well). The key distinction IPUMS harmonizes around is access to piped water, and secondarily whether distinctions can be made regarding exclusive use and the location of the spout on the property. The codes page for Water Supply is shown in Figure 9.

Figure 9. Water Supply Variable Codes Page (partial)



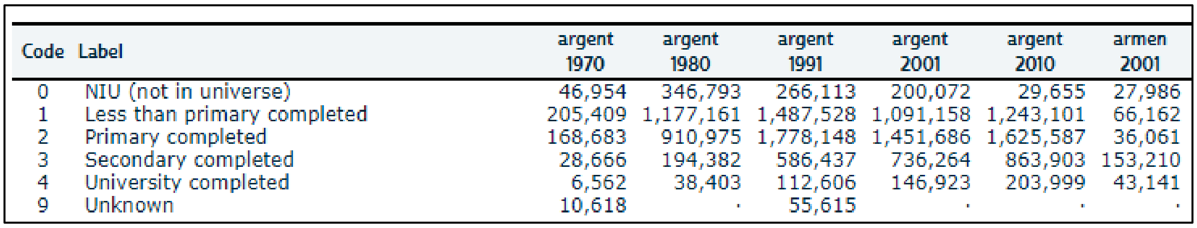
An 'X' indicates the category is available for that sample

Code	Label	AR	AR	AR	AR	AM	AM	AT	AT	AT	BY	BY	BO	BO
		1980	1991	2001	2010	2001	2011	1981	1991	2001	1999	2009	1976	1992
00	NIU (not in universe)	X	X	X	·	·	·	X	X	X	·	X	X	X
10	PIPED WATER	·	X	X	·	·	·	X	X	X	X	X	·	·
11	Piped inside dwelling	X	·	·	X	X	X	·	·	·	·	·	X	X
12	Piped, exclusively to this household	·	·	·	·	·	·	·	·	·	·	·	·	·
13	Piped, shared with other households	·	·	·	·	·	·	·	·	·	·	·	·	·
14	Piped outside the dwelling	·	·	·	·	·	·	·	·	·	·	·	·	·
15	Piped outside dwelling, in building	·	·	·	·	X	X	·	·	·	·	·	·	·
16	Piped within the building or plot of land	X	·	·	X	·	·	·	·	·	·	·	X	X
17	Piped outside the building or lot	X	·	·	·	·	·	·	·	·	·	·	X	X
18	Have access to public piped water	·	·	·	·	X	X	·	·	·	·	·	·	·
20	NO PIPED WATER	X	X	X	X	X	X	X	X	X	X	X	X	X
99	UNKNOWN	·	X	·	·	·	X	·	·	·	X	X	·	·

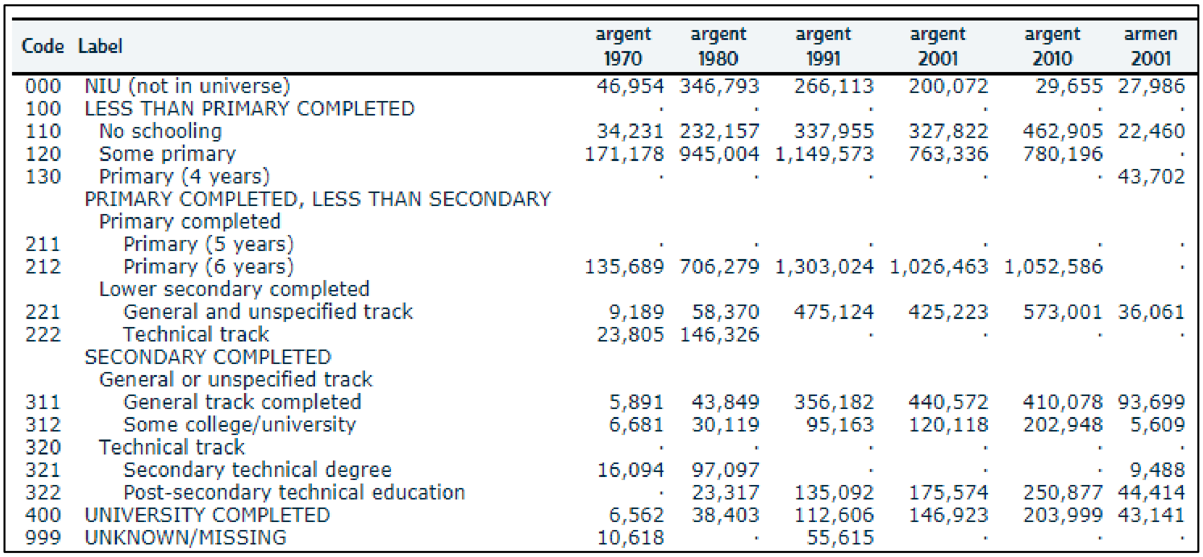
It was not possible within a single variable to accommodate all the concepts in water supply. And it is not indicated in most datasets when "piped" indicates clean water. In future, we intend to create another variable on the ultimate source of the water, for those samples that offer that detail, and perhaps we will be able to identify "clean" water in samples that will support that distinction.

Complex variables. IPUMS takes a different approach with key education and work variables: we coerce them into a classification intended to roughly follow international standards. The shoe-horning of categories into major groups can be uneven, and much detail in the original samples is sacrificed. The product is a simple, fairly consistent variable, but with a degree of noise. For the variables discussed above, we largely concede to the empirical reality of the categories we are presented with, and we fashion our harmonized classification in reaction to that, with some consideration of existing census standards and recommendations. With work and education we are much more aggressive. We are motivated to do so because few censuses provide income information; thus, education and occupation are the key socioeconomic status indicators typically available. They are critical control variables for many kinds of analyses. For education, we identify primary, secondary, and tertiary level completion. We roughly aim to identify people with 6 to 8, 11 to 12, or 15 to 16 years of education. The organization is broadly reflective of the ISCED 1997 classification (United Nations 1997). Figure 10 presents the 1-digit and 3-digit versions of Educational Attainment while displaying case-counts for each sample.

Figure 10. Educational Attainment Codes Pages: General and Detailed Versions



Code	Label	argent 1970	argent 1980	argent 1991	argent 2001	argent 2010	armen 2001
0	NIU (not in universe)	46,954	346,793	266,113	200,072	29,655	27,986
1	Less than primary completed	205,409	1,177,161	1,487,528	1,091,158	1,243,101	66,162
2	Primary completed	168,683	910,975	1,778,148	1,451,686	1,625,587	36,061
3	Secondary completed	28,666	194,382	586,437	736,264	863,903	153,210
4	University completed	6,562	38,403	112,606	146,923	203,999	43,141
9	Unknown	10,618	·	55,615	·	·	·



Code	Label	argent 1970	argent 1980	argent 1991	argent 2001	argent 2010	armen 2001
000	NIU (not in universe)	46,954	346,793	266,113	200,072	29,655	27,986
100	LESS THAN PRIMARY COMPLETED	·	·	·	·	·	·
110	No schooling	34,231	232,157	337,955	327,822	462,905	22,460
120	Some primary	171,178	945,004	1,149,573	763,336	780,196	·
130	Primary (4 years)	·	·	·	·	·	43,702
	PRIMARY COMPLETED, LESS THAN SECONDARY						
	Primary completed						
211	Primary (5 years)	·	·	·	·	·	·
212	Primary (6 years)	135,689	706,279	1,303,024	1,026,463	1,052,586	·
	Lower secondary completed						
221	General and unspecified track	9,189	58,370	475,124	425,223	573,001	36,061
222	Technical track	23,805	146,326	·	·	·	·
	SECONDARY COMPLETED						
	General or unspecified track						
311	General track completed	5,891	43,849	356,182	440,572	410,078	93,699
312	Some college/university	6,681	30,119	95,163	120,118	202,948	5,609
320	Technical track	·	·	·	·	·	·
321	Secondary technical degree	16,094	97,097	·	·	·	9,488
322	Post-secondary technical education	·	23,317	135,092	175,574	250,877	44,414
400	UNIVERSITY COMPLETED	6,562	38,403	112,606	146,923	203,999	43,141
999	UNKNOWN/MISSING	10,618	·	55,615	·	·	·

The internationally harmonized work and education variables lose much detail and are sometimes an imperfect fit for the national systems. This can be problematic where education is the dependent or key explanatory variable in a researcher's analysis. In recognition of education's importance, we create a separate harmonized variable for each country that is true to its specific education system. No attempt is made to apply a standard, only to harmonize around the classifications the country provides. This is harder than it sounds, as most countries have undergone changes or even complete reorganizations of their systems over the decades covered by IPUMS. Each country is therefore its own harmonization puzzle writ small, often requiring a good deal of research. Not surprisingly, educational attainment is one of the subjects upon which users most often provide feedback or identify errors.

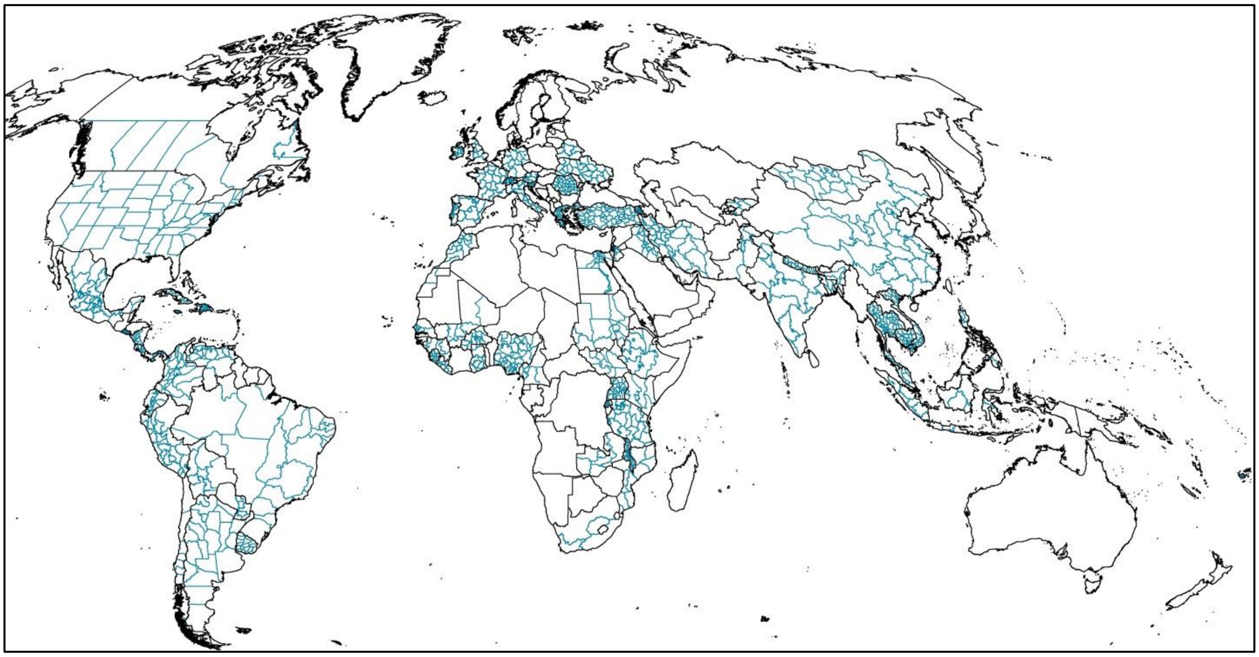
For occupation, we collapse the typical 100 to 300 categories in the original samples into a 9-category variable intended to mimic the major groupings in the 1988 ISCO standard as closely as possible (International Labour Office 2012). We do something similar in mapping industry using ISIC as a general guide (United Nations 2002). Due to its importance, we take an additional step with occupation. The ISCO occupation classification is used by many countries, and it can provide fully comparable detailed occupation data for all the countries that subscribe to it. Historically, these were more often developing countries, but in recent censuses developed countries are using it as well. ISCO has undergone several iterations. We make harmonized variables for the critical mass of samples providing 3-digit detail in both the ISCO-1968 and ISCO-1988 classifications, which are available for 27 and 57 samples, respectively.

For occupation and industry, we also make the full original classifications available through single cross-national variables that do not actually harmonize the codes. Thus, for the OCC (Occupation) variable, the codes for one sample mean entirely different things from another. The data are all organized into one place for user convenience, but there are no value labels with the data (they are available online).

Numeric variables. Numeric variables are sometimes continuous and in other cases are coded into intervals. It is always possible to recode continuous variables into intervals, but there is typically no way to perfectly harmonize one intervalled variable with another. Our preferred strategy with data provided as intervals is to code to the mid-points, creating a pseudo-continuous variable that can accommodate both grouped and ungrouped data.

Geography. Geography poses a unique set of issues for harmonization. Most countries have undergone changes in their administrative units over the past several decades, through merging, splitting, or moving a boundary. The goal of IPUMS is to harmonize subnational units spatially, so a province or district has the same spatial footprint in all time periods. This requires GIS boundary files, and IPUMS has created them from paper maps in all cases where digital versions were not available. This is costly in time and resources, and we offer the boundaries as a public good as well as using them in our processing. The process of harmonization requires overlaying each census's boundaries on each other and combining units as necessary to create entities that contain all the changes for an area within them. Figure 11 shows the first-level (largest) harmonized subnational units for each IPUMS country. Smaller second-level units are available for most countries as well. A researcher using these harmonized geographies knows they are holding space constant as they examine the attributes of the people and dwellings within those spaces. Spatial harmonization is essentially a least-common-denominator approach: if two units are combined in one census, they are combined in all of them. Detail is sacrificed to the goal of comparability. Since many researchers need the geography of the specific time and place of their study, IPUMS also provides the unaltered original geography for each census.

Figure 11. Harmonized Geography: 1st Administrative Level



Source variables

Harmonized variables are time-consuming to create and it is often difficult to prevent loss of information in the process. IPUMS has several hundred harmonized variables, but it cannot harmonize everything. To ensure that all the information in the census samples is accessible to researchers, IPUMS also provides the unharmonized source variables. These are the documented and labelled "standardized" variables discussed earlier, where no serious recoding was attempted except to deal with stray values or to minimally rationalize the codes. The source variables cover all the topics in the original censuses that it was not realistic to harmonize, because they are present in too few samples or are idiosyncratic in some way. The variables that serve as inputs to each harmonized variable are also identified. Researchers can therefore request the internationally harmonized version of employment status and each of the sample-specific variables used to construct it. This enables a motivated researcher to confirm our recodes or to devise their own harmonized version using only a subset of samples. There are over 30,000 unique source variables in the database providing access to the full detail of the original censuses.

Data use

Fifteen thousand registered IPUMS users from all over the world have created over 80,000 data extracts. Most of those extracts combine data from more than one census sample, and many include multiple countries. Such data pooling is only possible because the variables are harmonized. The most frequently accessed variables include many employment and education variables, which pose some of the most difficult harmonization challenges. A long tail of other variables requested by users includes all those discussed above. In each case, without IPUMS, users would be forced to reconcile codes in their own ad hoc way, which is hard to replicate and is error-prone. There is clearly great demand in the research community for harmonized data, and the most efficient use of scarce scientific resources is for specialized organizations to carry out this work and share it broadly with others.

Summary

In the final analysis, harmonization involves cost-benefit analysis. The goal is to make comparative research easier to conduct without obscuring the complications and thereby encouraging errors. Part of the job involves predicting how researchers are likely to use the data. Harmonizers must therefore have some subject matter expertise to strategize solutions effectively. But researchers are endlessly inventive, and a multi-purpose database will inevitably be used in ways that we cannot anticipate. Thus, a degree of conservatism is warranted, while providing enough documentation to allow users to exercise informed judgement.

An unfortunate reality of internationally harmonized data is the burden it places on the user. Both variable availability and the categories within those variables differ across samples. Using the most generalized versions of compositely-coded variables resolves many comparative issues, but certain definitional or population universe issues can still persist. And the composite-coding approach is not applicable to all variables. In sum, researchers are obligated to pay more attention to the metadata than they may be accustomed to, and it tends to be more complex. An ongoing challenge of our web dissemination system is to find better ways to convey the most important information without overwhelming users with details until they need them.

IPUMS is committed to harmonizing without losing information, but we see a role for least-common-denominator variables, and intend to develop them in the future. These will only offer categories that are fully comparable across all samples, and they will apply the most restrictive universe of people who answered the question among the available samples. In essence, the least detailed sample and the sample with the most restrictive universe will dictate the nature of these simplified harmonized variables. The main impetus from our perspective is the utility of such variables in our online tabulator. Researchers can tabulate millions of records in seconds using our online system, but recoding data or imposing case selection takes additional steps that many users would prefer to avoid. We also expect many users who download data will employ these highly-comparable simplified variables as controls in their models.

From our perspective, international population data harmonization is a puzzle whose subtleties are mostly amenable to human problem-solving rather than automation. But automation helps, and there is always more scope for it. At some point the costs come to outweigh the benefits, but that boundary will continue to shift in future as machine learning and other data science tools improve. We have already developed many utility programs that take advantage of semantic and coding similarities among data collections that are more coherent than the international censuses. For the foreseeable future, however, population microdata harmonization is bound to retain a significant component of human judgement.

References

Esteve A. and Sobek M. 2003. "Challenges and Methods of International Census Harmonization." Historical Methods 36: 66-79.

International Labour Office. 2012. *International Standard Classification of Occupations: Structure, Group Definitions and Correspondence Tables*. Geneva.

Madans J., Loeb M, and Altman B. 2011. "Measuring Disability and Monitoring the UN Convention on the Rights of Persons with Disabilities: The Work of the Washington Group on Disability Statistics." *BMC Public Health* 11, Supplement 4.

Ruggles S. et al. 2015. "The IPUMS Collaboration: Integrating and Disseminating the World's Population Microdata." *Journal of Demographic Economics* 81: 203-216.

Sobek M. 2016. "Data Prospects: IPUMS-International." In Michael White, ed., *International Handbook of Migration and Population Distribution*. New York: Springer, 2016, 157-174.

Sobek M. et al. 2011. "Big Data: Large-Scale Historical Infrastructure from the Minnesota Population Center." *Historical Methods* 44: 61-68.

United Nations. 1997. *International Standard Classification of Education: ISCED 1997*. United Nations Educational, Scientific and Cultural Organization. New York. [http://www.unesco.org/education/information/nfsunesco/doc/isced_1997.htm]

United Nations. 2002. *International Standard Industrial Classification of All Economic Activities (ISIC), Revision 3.1*. Department of Economic and Social Affairs, Statistics Division. New York.

United Nations. 2007. *Principles and Recommendations for Population and Housing Censuses (Revision 2)*. Department of Economic and Social Affairs, Statistics Division. New York.