Challenges and Methods of International Census Harmonization

Albert Esteve and Matthew Sobek
Minnesota Population Center

The development of IPUMS-International (Integrated Public Use Microdata Series – International) involves working with census materials of different vintages and institutional origins. The samples are of varying quality and employe different formats and variable coding schemes. This article describes the procedures we developed to cope with the range of variations encountered. The focus is on the challenges of harmonization: the creation of a unified, consistent data series from these disparate census samples. The discussion is based largely on our experience with the first round of countries incorporated in the database for release in Spring 2002: Colombia, France, Kenya, Mexico, the United States, and Vietnam.

We harmonized the international microdata samples in two distinct stages. The first stage consisted of standardizing the varying formats of the original microdata and correcting any errors uncovered in the process. The second stage involved harmonizing each variable across all samples, including determining the availability of comparable variables, compiling the existing information on each, and designing the variable coding schemes and corresponding documentation.

Format Standardization

Census microdata exist in a surprisingly wide range of data structures and file formats. The oldest datasets—those dating from the 1960s and 1970s—are often plagued by internal structural inconsistencies, a byproduct of the severe constraints on computing and data storage in those decades. Even the most recent samples, however, required substantial effort to verify that they were free of data format problems. Such errors typically affected only a small fraction of cases; nevertheless, these problems had to be addressed systematically to produce clean sample data.

The raw data files were preserved in a variety of formats. The simplest files were rectangular, with geographic, dwelling, household, and family information replicated on each person record. More complex file structures included multiple nested record types in a single file, records stored in separate files that had to be linked together, and separate files with different record layouts for various segments of the population.

We began by reformatting each sample into a hierarchical format consisting of a household record followed by person records for each individual in the household. Any geographic or dwelling-level information was replicated on each respective household record. This data reformatting produced a standardized input structure for subsequent recoding routines. Just as important, the data

manipulation often exposed problems that could not be identified from a detailed examination of data frequencies or cross-tabulations. Thus, the process of restructuring the data was an integral aspect of diagnosis and cleaning.

Varieties of Data Structures

The following examples describe the most common data structures and how they were treated:

- Multi-level hierarchical records. The data structures of the most complex samples had as many as four nested record types identifying the starting points of each geographic area, dwelling, and household. In such cases, we collapsed all levels of detail above the person into a single household record. With such a detailed hierarchy in the original data, any irregularity in the sequence of record types creates widespread data problems. In the case of Vietnam 1989, for example, missing dwelling records sometimes forced us to infer the breaks between multi-household dwellings and stand-alone households.[1]

- Dwelling-level samples. Some samples had dwelling records but lacked household records. Sometimes the distinction was essentially terminological, and for practical purposes the dwelling records could be considered household records (e.g., Colombia 1985). In other samples, there could be multiple distinct households within dwellings, each with its own head (e.g., Mexico 2000). In these instances we constructed household records by duplicating the dwelling information for each household in the unit, while eliminating the dwelling record itself. Any information unique to specific households within the dwelling was taken from the household head's record.[2]

- Rectangular format. In rectangular files each record is an individual, and any household, family, and geographic information is repeated on every record. Rectangular format is the preferred output for most data users, but it is inefficient to store the information this way. For those samples that were provided in this format (e.g., Mexico 1990), we removed the household-level information from each person and constructed a single household record for each unit using the data on the household head's record. When both dwelling and household information was embedded in the rectangular format, the divisions between units were logically identified and separate household records were created for each household.

[1] By design, part of the Vietnam 1989 sample lacked dwelling records. Approximately half of all households were not asked the dwelling questions. In the sample it was not always clear where, within a given enumeration area, the last dwelling ended and the non-dwelling households began. Strict interpretation of the data structure would have suggested dwellings with hundreds of households of perhaps 20 square meters in area.

[2] In accordance with our practice of never losing meaningful information, we created a household sequence number that gives each household a unique identification number within the dwelling. Researchers can recombine the households into dwellings using this ID.

- Individual (not household) samples. Some files are samples of individuals rather than households (e.g, Colombia 1964). This limits their utility for many purposes, but the microdata still provide far more information than published tabulations. To standardize these samples, we took geographic information and any other household-type data off the person record and constructed a "household" record from it. Every person in such samples has their own household record.

- Group quarters individuals. Many samples lacked household records for persons in institutions or other group quarters situations. We constructed household records for each group quarters individual when it was clear they were sampled individually. If the data indicated that whole groups of persons were sampled from a specific group quarters (e.g, Kenya 1999), we kept those units intact by inserting a single household record preceding each group.

- Censuses with multiple record layouts. It is not unusual for censuses to use separate questionnaires for different subpopulations (e.g., institutions, collective households, migrants, homeless persons, indigenous peoples). In some cases the different questionnaires are reflected in the microdata as separate files with varying column formats and variable availability. To avoid a proliferation of record layouts, we reformatted such files to match the record layout of private households in each census.

- Separate files requiring matching. Some censuses are organized into multiple record types stored in separate files designed to be linked together by means of a common identification number. These record types can include mortality, fertility, and group quarters records as well as person, household and dwelling records. For such samples, small imperfections in the data structures can cause significant problems. If there was a problem matching records, we used the following information to perform the match: variables common to both household and person files, sequence in the original file within geographic unit, and direct comparison of records with non-unique identification numbers for the best potential match.

Samples from Full-Count Data

Experience has taught us that national statistical offices do not always verify the consistency of different hierarchical levels within census data. We sometimes uncovered mismatches between dwellings, households, and persons. The marginal distributions of both individual and household characteristics generally were sound, but inconsistencies between record types created problems for the construction of microdata samples. These included households without persons, persons without households, or households blended together. Such overt data problems rarely involved large numbers of cases; nevertheless, they had to be addressed to produce clean and consistent datasets.

Space constraints prevent us from describing here the full variety of data problems we encountered and explaining our solutions. Each sample was different, and we employed whatever internal data were available to arrive at a strategy for logical or

probabilistic correction of errors. One particular example, however, is worth describing in some detail, because it demonstrates the possibilities for correcting data problems when full-count data are available from which to draw a sample.

For the Colombian census of 1973, we began with the 100 percent population microdata used to create published tabulations. The data were in separate household and person files. Attempts at matching the files by household identification (ID) number uncovered an array of data errors. In the household file, some households shared the same ID; others had corrupted data as part of the ID. In the person file, there were separate distinct blocks of persons with the same ID, sets of two households of persons blended together, households of persons split up by intervening households, and other irregularities. To construct a clean sample of the Colombian census, we used a sequence of diagnostic procedures to mark records in the household file that exhibited any of these format errors. In the end, we classified 2.9 percent of the household records as bad (i.e, having some sort of problem, however minor).

We then drew a 10 percent sample of household records. After a random start within geographic units, we marked every tenth household in the original data. If the 10th household was flagged as bad, we substituted the most proximate household with the same value for the "number of persons" variable. This procedure is essentially the same as the hot deck allocation method used by the U.S. Census Bureau to infer characteristics of non-responding households. The resulting sample of household records matched cleanly to the person file. By identifying donor households in close geographic proximity to the corrupted households, we were able to maintain representativeness. There are no detectable systematic biases in the completed 10 percent sample; on all characteristics, the sample falls within the expected confidence intervals when compared to the complete count.

We expect to use this procedure in the future in cases where there are significant data integrity issues and we have complete-count or high-density data from which to draw a sample of lesser density. Such situations are more likely to arise than we imagined when we began the project. In many cases old full-count data survive in archives around the world, but the country in question never created a sample. In Latin America, we have found widespread willingness to allow us to draw public-use samples from full-count data, and we expect other such opportunities to arise in the future.

Variable Harmonization

Most of the IPUMS-International work process centered on variable-level harmonization. The goal was to create variables that are consistent across time and space, enabling users to carry out cross-national research with a temporal dimension. While the dataset reformatting and cleaning described above treated each sample in isolation, variable harmonization required consideration of all of the datasets simultaneously. Virtually all variable-level harmonization is imperfect, because of variations in the wording of questions, the classifications employed by each census, and the cultural meanings of census concepts. For this reason, a major component of variable harmonization is identification and documentation of potential incompatibilities. Important differences

between variable categories across samples should be made sufficiently clear to enable intelligent use of the data. We therefore have two tasks: we must both recode the data for maximum compatibility and make evident to users any ambiguities or compromises of the resulting classification.

There are three components to the process of variable harmonization: compiling the existing documentation, determining variable availability, and designing the harmonized coding scheme and associated documentation. They are discussed in turn below, followed by an example of variable harmonization.

Documentation Resources

Sample documentation played an essential role in the overall process of integration, and was relevant to all stages of variable harmonization including determining availability, designing coding structures, and writing the variable discussion for the web site. Enumeration forms, enumerator instructions, and microdata sample codebooks were the main sources of variable documentation. One of the basic tasks of the project was to draw together this information, which was not always available for older censuses from the national statistical agencies. Fortunately, the United Nations Statistics Division and the UN Demographic Center for Latin America and the Caribbean (CELADE) were able to provide key documents that we lacked.

The international dimension of the database required careful attention to differing meanings of questions and responses, and involved comparing sometimes strikingly different classification systems. To complicate matters, the quality and quantity of variable documentation varied considerably across samples. Codebooks supply the basic information needed to read a file, and they usually provide variable names and a label for all possible values for each variable. Sometimes the sole source of information is a data definition file for a statistical package, or a frequency distribution with labels generated by a statistical package, sources that are at least one iteration removed from the original documentation. The existing codebooks were sometimes incomplete or inaccurate, with some values and even entire variables left undocumented. As with most data problems, this was more common in older datasets. In some cases it was possible to identify an undocumented variable because the microdata format exactly mimicked the order of the questions on the census questionnaire. In other cases, we have been unable to discover the meaning of particular columns of raw microdata.

In addition to codebooks, we always had access to the original census questionnaires, which show the census wording on the forms and any pre-defined categories for responses. Undocumented values for known variables could sometimes be inferred by reference to the responses printed on the forms. For most censuses, we also had the instructions to the enumerators. And in some cases we had post-enumeration processing instructions that detailed how the data from the manuscripts were transformed into the microdata.

To supplement such pre-existing sample documentation, IPUMS-International commissioned a series of topical essays from specialists in the respective countries. These were designed to provide insight on integration problems and possibilities from the perspective of persons with extensive experience with a country's census data. Typically, these essays were written on clusters of variables, such as economic characteristics, education, or housing. The value of these contributions varied, but in some cases they provided the only information at our disposal on certain variables beyond what was printed on the census form itself. The essays also allowed the specialists to highlight what they considered to be major comparability or data quality issues, drawing from their own knowledge and experience.

Variable Availability

Before the variables could be harmonized, we needed to determine the availability of variables across all samples. With hundreds of variables in numerous original languages, this was not a trivial task. Moreover, language differences aside, equating seemingly similar variables was not always straightforward because of varying conceptual and terminological conventions. Despite United Nations and other international influences, even comparable variables are referred to differently across countries.[3] Inevitably, there were gray areas where variable content did not precisely overlap. Equating similar, but not identical, variables from two samples required weighing convenience for users against encouraging specious comparisons.

Fortunately, most censuses around the world share a core set of questions that are conceptually comparable. These include the basic demographic characteristics of the population and often extend to economic activity and education. The categories may differ, but there is no doubt in most instances that these are essentially the same variables. The similarity among international censuses is not accidental. The United Nations has been influential in standardizing census practices among many countries, particularly those with a short tradition of census-taking. The UN has developed guidelines concerning what questions to ask and how best to classify the results (United Nations 1998a, 1998b). To the extent that multiple countries have adopted these standards, it has reduced variations in both subject content and terminology. The UN guidelines also offered us a convenient standard for grouping and naming variables.

[3] For example, the "class of worker" variable in the United States (i.e., employer, self-employed, employee) is referred to as "status in employment" in countries that subscribe to the UN census terminology. "Status in employment" under UN terminology is unrelated to the U.S. census concept "employment status" (labor force status and unemployment); under UN terminology the latter topics are covered in "activity status."

Table 1. Selected Subject Area Availability, by Country and Census Year

	Colombia				France					Kenya		Mexico				United States				Vietnam	
	64	73	85	93	62	68	75	82	90	89	99	60	70	90	00	60	70	80	90	89	99
Geography and internal migration																					
Place of usual residence	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Place of birth	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	.	.
Duration of residence	x	x	.	.	.	.	.	.	.	.	x	x	x	.	.	x	x	x	x	.	.
Place of previous residence	x	x	.	.	.	.	.	.	.	.	.	x	x	.	.	.	.	.	.	.	.
Place of residence at a specified date in the past	.	.	x	x	x	x	x	x	x	x	x	.	.	x	x	x	x	x	x	x	x
Household and family structure																					
Relationship to head of household/householder	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Demographic and social																					
Sex	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Age	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Marital Status	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Citizenship	.	.	.	.	x	x	x	x	x	x	x	x	.	.	.	.	x	x	x	.	.
Religion	.	.	.	.	.	.	.	.	.	.	x	x	x	x	x	.	.	.	.	.	x
Language	.	.	.	.	.	.	.	.	.	.	.	.	x	x	x	.	.	x	x	.	.
National and/or ethnic group	.	.	.	x	.	.	.	.	.	x	x	x	.	.	x	x	x	x	x	x	x
Fertility and mortality																					
Children ever born	.	x	x	x	.	.	.	.	.	x	x	x	x	x	x	x	x	x	x	x	x
Children living	.	x	x	x	.	.	.	.	.	x	x	.	.	.	x	.	.	.	.	x	x
Date of birth of last child born alive	.	x	.	x	.	.	.	.	.	x	x	.	.	.	x	.	.	.	.	x	x
Deaths in the past 12 months	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x
Maternal or paternal orphanhood	.	.	.	.	.	.	.	.	.	x	x	.	.	.	.	.	.	.	.	.	.
Age, date or duration of first marriage	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	x	x	.	.	.
Education																					
Literacy	x	x	x	x	.	.	.	.	.	x	.	x	x	x	x	.	.	.	.	x	x
School attendance	.	x	x	x	.	.	.	.	.	x	x	.	x	x	x	x	x	x	x	x	x
Educational attainment	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Field of education and educational qualification	.	.	.	.	x	x	.	.	.	.	.	.	.	.	x	.	.	.	.	x	x
Economics																					
Employment status	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Time worked	x	.	x	.	.	.	.	.	.	.	.	x	x	x	x	x	x	x	x	.	.
Occupation	x	x	.		x	x	x	x	x	x	.	x	x	x	x	x	x	x	x	x	x
Industry	x	x	.	x	x	x	x	x	x	.	.	x	x	x	x	x	x	x	x	x	x
Class of worker	x	x	x	x	x	x	x	x	x	x	.	x	x	x	x	x	x	x	x	.	.
Income	.	.	.	.	.	.	.	.	.	.	.	.	x	x	x	x	x	x	x	.	.
Institutional sector of employment	.	.	.	.	x	x	x	x	x	.	.	.	.	.	.	.	.	.	.	.	x

Place of work	.	.	.	.	x	x	x	x	x	.	.	.	.	.	x	x	x	x	x	.	.	
International migration																						
Country of birth	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	.	.
Citizenship	.	.	.	.	x	x	x	x	x	.	.	.	.	.	.	x	x	x	x	x	.	.
Year or period of arrival	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	x	x	x	x	.	.
Disability																						
Disability	.	.	.	x	x	.	.	.	.	.	.	.	.	.	x	.	x	x	x	x	.	.
Cause of disability	.	.	.	.	x	.	.	.	.	.	.	.	.	.	x	.	.	.	.	.	.	.

Notes: Samples are identified by the last two-digits of their census year. An "x" indicates the topic is available in that sample.

Table 1 presents a partial listing of topical coverage of variables among the six countries in the 2002 IPUMS-International data release. This preliminary release includes a subset of variables that appear in the censuses. The subject areas are organized according to the UN categorization of topics, and a particular subject in the table may actually represent multiple variables on that topic. Variable availability differs by country and year but, in general, national differences are more important than variations within countries over time.

Even when our work on these censuses is complete, the microdata samples will not always contain variables corresponding to every item on the census questionnaires. Confidentiality concerns are one cause of discrepancies, and are most evident in limited geographic detail. Each country imposes different standards concerning the smallest geographic unit identifiable in the microdata. The smallest geographic unit ranges from places with 20,000 inhabitants in some countries to entire regions in others. Other variables considered too sensitive or too great a risk to privacy may also be left out of the samples. In addition, particular variables may be coded in such a way that they do not retain all of the responses that were included on the census forms—typically to prevent the identification of small population subgroups. A national statistical office may also reject the inclusion of a variable because the census question was deemed flawed and the data unreliable. Finally, some variables may have been left out of the microdata because there were insufficient resources to process the data, such as translating handwritten responses into a numeric classification for the computerized data.

Harmonization Process

The international census samples use differing numeric classification systems for their variables. Harmonization of these codes was a central aspect of the project. Variable design often influences the analytical strategies adopted by researchers, and coding schemes therefore had to be developed with care. The aim was to create a comparable set of codes for each variable that meant the same thing across countries and over time.

After ascertaining the availability of a given variable, we compiled the entire range of codes and labels associated with it in all datasets. Except for the simplest variables, the labels by themselves were usually insufficient to categorize the values in the context of all the countries in the database. The primary source for interpreting the meaning of the variable categories was the original census questionnaire. Questionnaires, however, often give little information beyond the universe and specific names of the pre-defined responses. When the questionnaire was insufficient, the census enumerators' instructions usually provided the necessary information to discern the meaning of particular codes. When there was conflicting or ambiguous information, we considered the questionnaire to preempt other sources, because it was most immediately in mind when the census questions were being answered. Lastly, we had recourse to the materials and expertise of our international partners in interpreting the codes. When the data contained an undocumented value that could not otherwise be interpreted, we coded it as missing and allocated it through probabilistic or logical procedures.

The coding design attempted to balance two competing goals. First, we wanted to keep the variables simple and easy to use for comparisons across time and space. This required providing the lowest common denominator of detail that was fully comparable. At the same time, however, we were committed to retaining all meaningful detail in each sample, even when it was unique to a single dataset. We employed several approaches to achieve these competing goals. Some variables were compatible and could simply be recoded into an internationally consistent classification with no loss of detail. Most variables were more complex, however, and required the use of a composite coding scheme to encompass the full range of differences. In the composite system, the first digit or digits of a variable describe categories that are comparable across all datasets. Subsequent digits provide details that are available in some samples but not others.

The classification scheme for marital status illustrates the composite coding approach. Under the IPUMS-International design, shown in Table 2, the first digit of marital status has four categories: single, married/in-union, separated/divorced/spouse-absent, and widowed. This is the maximum number of categories consistently distinguishable across all samples. The distinction between divorced and separated is not maintained in all censuses; therefore, these two categories are combined in the fully comparable first digit of the marital status variable. With the second digit, divorced and separated persons can be distinguished, as can formal marriages from consensual unions. The third and final digit differentiates among types of marriages (e.g., civil, religious, and polygamous) using information only available for select countries.

Table 2. Coding Scheme and Category Availability for Marital Status

		Colombia				France					Kenya		Mexico				United States				Vietnam	
Code	**Label**	**64**	**73**	**85**	**93**	**62**	**68**	**75**	**82**	**90**	**89**	**99**	**60**	**70**	**90**	**00**	**60**	**70**	**80**	**90**	**89**	**99**
100	SINGLE/NEVER MARRIED	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	MARRIED/IN UNION																					
210	Married (not specified)	x	x	x	x	x	x	x	x	x	x	x	.	.	.	.	x	x	x	x	x	x
211	Civil	.	.	.	.	.	.	.	.	.	.	.	x	x	x	x	.	.	.	.	.	.
212	Religious	.	.	.	.	.	.	.	.	.	.	.	x	x	x	x	.	.	.	.	.	.
213	Civil and religious	.	.	.	.	.	.	.	.	.	.	.	x	x	x	x	.	.	.	.	.	.
214	Polygamous	.	.	.	.	.	.	.	.	.	x	x	.	.	.	.	.	.	.	.	.	.
220	Consensual union	x	x	x	x	.	.	.	.	.	.	.	x	x	x	x	.	.	.	.	.	.
	SEPARATED/DIVORCED/SPOUSE ABSENT																					
310	Separated or divorced	.	x	x	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
320	Separated	x	.	.	.	.	.	.	.	.	x	x	.	x	x	x	x	x	x	x	x	x
330	Divorced	.	.	.	.	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
340	Married, spouse absent (n.s.)	x	x	x	x	x	x	x	x	x	x	x	.	.	.	.	x	x	x	x	x	x
341	MSA, civil	.	.	.	.	.	.	.	.	.	.	.	x	x	x	x	.	.	.	.	.	.
342	MSA, religious	.	.	.	.	.	.	.	.	.	.	.	x	x	x	x	.	.	.	.	.	.
343	MSA, civil and religious	.	.	.	.	.	.	.	.	.	.	.	x	x	x	x	.	.	.	.	.	.
344	MSA, polygamous	.	.	.	.	.	.	.	.	.	x	x	.	.	.	.	.	.	.	.	.	.
350	Consensual union, spouse absent	x	x	x	x	.	.	.	.	.	.	.	x	x	x	x	.	.	.	.	.	.
400	WIDOWED	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Notes: Samples are identified by the last two-digits of their census year. An "X" indicates the category is available in that sample.

To make the harmonization process manageable, for most variables we separately performed a preliminary integration of each country's samples before turning to the final international integration. This focused the compilation of information first on the time

dimension, highlighting changes in national census practices that might have been submerged in the wider international context. The integration essays by international collaborators were particularly useful at this stage.

We then considered all countries simultaneously to create an integrated international variable coding scheme. We were forced to develop most of the coding schemes ourselves. United Nations census recommendations were the closest thing to a standard, but they were meant as a guide to census-taking, not as a primer on integrating already-existing samples.[4]

Under the composite coding approach, our intent was to make the first one or two digits comparable across all samples and add trailing detailed digits to retain variations available in some samples and not others. In contrast to our experience with IPUMS-USA, there were instances in which even at the most general level we did not make the categories absolutely comparable. For example, for the relationship variable, the 1999 census of Vietnam combined "other relatives and nonrelatives" in a single category, whereas every other sample made this critical distinction. Full comparability in the first digit was achievable, but only at the cost of making the codes much more cumbersome for the majority of samples, requiring users to go past the first digit of the variable to get this fundamental kinship information. In essence, we allowed some incompatible coding to persist in order not to inconvenience the majority of users with a clumsy coding structure. The price of increased usability is the necessity for greater vigilance by the users. In some instances—such as occupation—it was impractical to develop a compatible composite coding scheme without losing information. In these instances, we provided a separate unrecoded version of the variable to ensure to ensure that no original detail was sacrificed.

For each variable, we developed an integrated variable description and comparability discussion. This involved assessing incompatibilities and highlighting potential problems. With widely varying source data, even an elegant coding design inevitably requires supporting text on comparability issues not reflected in the category labels. The documentation also covers such matters as the population universe for each variable. We empirically verified the universe for each variable, and found many instances in which the universe in the data differs from the census questionnaires or the microdata codebooks.

One can never fully anticipate how census microdata will be used, so all relevant documentation must be made accessible. Saying too much, however, dilutes the most important comparability issues. To avoid overwhelming the user, more technical and mundane descriptive information is accessible from hypertext links on the main variable pages. These links include images of the original questionnaires and instructions.

Employment Status: An Example

[4] For some of the most difficult variables—educational attainment, occupation and industry—there were international classifications that served as useful frameworks for integration (International Labor Office 1990; United Nations 1990; UNESCO 1997).

Table 3. Code Availability for Employment Status, Selected Censuses

	Colombia 1993		**France 1975**		**Kenya 1999**
B	Missing	0	Inactifs autres que ceux cites ci dessous	B	N/A (under age 5)
1	Busca trabajo, tenia trabajo	1	Actifs ayant un emploi	0	[undocumented]
2	Busca trabajo, primera vez	3	Personnes sans emploi et en recherchant	1	Work pay profit
3	No trabajo pero tenia	4	Anciens actifs (retraites, retires des affaires…)	2	On sick leave
4	Trabajar	5	Etudiants et eleves nes avant le 1/1 1962	3	Family holding
5	Estudiar	6	Militaires du contingent	4	Family agricultural holding
6	Oficios hogar			5	Seeking work
7	Incapacitado			6	No work available
8	Jubilacion			7	Student
9	No response			8	Retired
0	Otra situacion			9	Incapacitated
				10	Home maker
				11	Other
	Mexico 2000		**United States**		**Vietnam 1989**
B	Missing	0	N/A	1	Worked 6 months and over
10	Trabajo	10	At work	2	Worked permanently less than 6 months
13	Trabaja se declara que busca trabajo	11	At work, public emergency	3	Worked temporarily less than 6 months
14	Trabaja se declara que es estudiante	12	Has job, not at work last week	4	Unemployed
15	Trabaja se dedica a quehaceres del hogar	13	Armed forces	5	Student
16	Trabaja se declara que es jubil o pens	14	Armed forces	6	Household duties
18	Trabaja se declara que no trabaja	15	Armed forces, not at work last week	7	Invalid
19	Trabaja no se tiene informacion	20	UNEMPLOYED	8	Others
20	Tenia trabajo per no trabajo	21	Unemployed, experience worker		
30	Busca trabajo	22	Unemployed, new worker		
40	Estudiante	30	NILF		
50	Quehaceres del hogar	31	Housework		
60	Es jubilado o pensionado	32	Unable to work		
70	Esta incapacitado permanente para trabajar	33	School		
80	No trabaja	34	Other		
99	No especificado				

The IPUMS-International variable "employment status" illustrates the overall process of harmonization. First, we determined the availability of the variable. Among the first-release countries, every sample had conceptually comparable information on employment status. The variables were called different things, and in some cases (e.g., Mexico 1960) the information had to be constructed from more than one variable in the original microdata. Once we determined the availability of employment status across datasets, we compiled from the different codebooks the specific variable categories that were identified in each sample. A partial listing of these categories is presented in Table 3.

The international harmonization began by determining the key distinctions maintained across countries and consequently the maximum number of sustainable categories that could be identified universally. For employment status, the critical information each census aimed to capture was participation in the labor force and unemployment. The resulting IPUMS-International classification for employment status is shown in Table 4. The leftmost columns give the IPUMS codes and their labels. The columns to the right show the corresponding original codes.[5]

Table 5. Translation Table for Employment Status

Harmonized Codes and Labels			Source Data Codes (selected samples)									
IPUMSI Code	IPUMSI Label	Census	Col 64	Col 93	Fra 62	Fra 75	Ken 99	Mex 70	Mex 00	US 60	Viet 89	Viet 99
0000	N/A		*,5	B	*	B	BB	0	BB	00	B	B,1
	ACTIVE (In Labor Force)											
1000	EMPLOYED, not specified		1								1	
1100	At work			4	1	1	01	1	10	10		
1101	At work, and 'student'								14			
1102	At work, and 'housework'								15			
1103	At work, and 'seeking work'								13			
1104	At work, and 'retired'								16			
1105	At work, and 'no work'								18			

[5] Although they perform the same function, the IPUMS-International translation tables are more sophisticated than those used in the creation of IPUMS-USA. The new tables can accommodate any number of datasets and can include columns for the frequency distribution, original-language labels, and English-language labels for each dataset (not shown in Table 5). The additional features are necessary because of the greater challenges of international integration and the large number of samples. With many datasets from many countries, we sometimes modified the coding structure for a given variable during development. Working with the codes alone in such an iterative process—as we did with IPUMS-USA—would have been impractical.

1106	At work, public emergency								11		
1107	At work, family holding, not specified										
1108	At work, family holding, not agricultural					03					
1109	At work, familiy holding, agricultural					04					
1110	Working and studying (France)										
1200	Have job, not at work last week		3			02		20	12		
1300	Armed forces								13		
1301	Armed forces, at work								14		
1302	Armed forces, not at work last week								15		
1303	Military trainee (France)			8	6						
2000	UNEMPLOYED, not specified	2			3	05	2	30	20		
2001	Unemployed (Vietnam)									4	5
2002	Worked <6 months, permanent job									2	
2003	Worked l< 6 months, temporary job									6	
2100	Unemployed, experience worker		1						21		
2101	Seeking work, worked less than 3 months			2							
2102	Seeking work, worked 3 to 6 months			3							
2103	Seeking work, worked 6 to 12 months			4							
2104	Seeking work, worked > 1 year			5							
2105	Seeking work, experience unspecified			6							
2200	Unemployed, new worker		2	7					22		
3000	INACTIVE (Not in Labor Force)								30		
3100	Housework	3	6			10	3	50	31	6	2
3200	Unable to work/disabled	7	7			09		70	32	7	4
3300	In school	4	5	9	5	07		40	33	5	3
3400	Retirees and living on rent	8						60			
3401	Living on rent payments										
3402	Retirees/pensioners		8		4	08					
3500	Elderly	6									
3600	No work available/discouraged					06					
3700	Inactive, other reasons	9	0	0	0	11	4	80	34		6
[alloc]	ALLOCATED VALUES (unknown/missing)		9			00	9	99			9

Note: In the source data columns: a comma indicates more than one code was coded to the respective IPUMS-International value; an asterisk means programming logic was used; B indicates a blank in the source data. Allocated values are assigned codes based on probabalistic editing procedures.

The first digit of employment status has three categories that are largely comparable across all samples: employed, unemployed, and not in the labor force. After the first digit, national and temporal variations become evident. Among the unemployed, some samples distinguished between persons with past work experience (experienced unemployed) and persons seeking work for the first time (new workers). The number of categories distinguished among the inactive population varies widely across samples.

The final step was writing the variable documentation. One key aspect of employment status that required explanation was unemployment. The unemployed population is difficult to define consistently across countries. We applied United Nations and International Labor Organization standards in defining the unemployed as persons who are out of work and actively seeking a job. Some countries have relatively small paid-labor sectors and irregular labor markets, and definitions of unemployment vary. For example, Kenya identified persons who were not working simply because no work was available, explicitly referring to them as "discouraged" workers in the 1999 enumeration instructions (Table 5). In Kenya these persons were considered unemployed, but in IPUMS-International they are coded as "not in the labor force" to maintain consistency with other countries.

A second major documentation issue concerns the varying reference period for the employment status question. For most samples, employment status was reported with respect to the day of the census or within a specified week prior to the census. For Vietnam, however, the reference period was the previous year—amounting to "usual employment status" over this period. These and other points are covered in the variable description and comparability discussion for employment status.

Conclusion

The integration of international census microdata is challenging. The number of samples and variables is large, and the data quality and source documentation are uneven. Because of cultural differences, the meaning of some variables and categories across countries is uncertain. Inevitably, we will miss some nuances of interpretation. If the experience of IPUMS-USA is a guide, however, our expert users will constitute an army of fact-checkers. IPUMS-International will remain a work-in-progress for the foreseeable future, and user feedback will help guide future revisions.

Despite all the difficulties, the effort to harmonize is important. If researchers were forced to reconcile the censuses themselves, they would inevitably make mistakes, different analyses would be incompatible with one another, and there would be wasteful duplication of effort. Most important, without harmonization few cross-national studies would be undertaken.

We anticipate many revisions and additions to IPUMS-International. Work on China and Brazil is proceeding, and we are preparing additional variables for each of the samples that is already included in our preliminary release. In June 2003, we will begin incorporating approximately seventy censuses from sixteen additional Latin American countries into the database. We are planning additional projects to add censuses from Europe, Asia, Africa, and the Pacific Islands. If these projects are successful, IPUMS-International will eventually become a comprehensive source of information on a crucial era of world economic and demographic history.

References

International Labor Office. 1990. International Standard Classification of Occupations (ISCO-88). Geneva.

UNESCO. 1997. The International Standard Classification of Education (ISCED 1997). Paris.

United Nations. 1990. International Standard Industrial Classification of All Economic Activities (ISIC-88). United Nations Statistics Division. Department of Economic and Social Affairs, New York.

United Nations. 1998a. Principles and Recommendations for Population and Housing Censuses. United Nations Statistics Division. Department of Economic and Social Affairs, New York.

United Nations. 1998b. Recommendations for the 2000 Censuses of Population and Housing in the ECE Region. United Nations Economic Commission for Europe and Statistical Office of the European Communities. Statistical Standards and Studies, No. 49. New York and Geneva.