IPUMS International Introduction Webinar
March 28, 2019 (11:00 a.m.-12:00 p.m. CT)

Questions and Answers

- **How does the IPUMS census collection compare to other compiled data collections at international organizations (e.g. World Bank, UN, etc.)?**
  The biggest difference is that those organizations do not disseminate individual-level data. The UN and the World Bank have valuable databases full of statistical summaries, constructed indicators, and aggregate data. IPUMSI disseminates individual-level data files, containing information asked in the census for each individual included in the sample. The IPUMS International data extract system allows the user to build customized, pooled datasets. IPUMS adds significant value for comparative research by creating an additional layer of harmonized variables and enabling pooling of data right in the data request system. IPUMS data are also discoverable through the World Bank microdata catalog. The World Bank has compiled metadata and information about accessing microdata, some of which is directly accessible through their tool, but all IPUMS files are discoverable there too, and World Bank provides a link directing the user to IPUMS.

- **How is getting this different from getting data from other microdata access sources?**
  IPUMS International is different from other microdata resources (e.g. DHS) because IPUMS data are harmonized. Behind the IPUMS website is a large database where all the data share a common structure and where a subset of harmonized variables are shared across all country years. Users can pull the variables and samples they need for their research. When a data user downloads a data set from IPUMSI, and opens it up on their computer, everything is already pooled into a single file. The data user does not have to do any renaming or recoding of variables, no appending or merging of files. For DHS, as an example of another microdata source, the user must download a single file for each survey, pool them together, and harmonize variables as needed. IPUMSI is beneficial because it makes data manipulation and data cleaning almost non-existent for the user.

- **Have you found many errors in the harmonized variables? I've heard that in the past there were issues with coding, and it has been a barrier for suggesting IPUMS to my peers.**
  We take pride in the harmonization work, and we have a number of programs and tools that help us carry out and check our work. However, decisions about what variables mean, what they mean across countries, and how to align the variables are made by human brains. So of course, as humans, we do make mistakes. We have a thorough review process where we try to catch those errors before the data gets to the website, but we do make errors. When users register for access to use the data, they agree to several items regarding good data stewardship including alerting us (at IPUMS) to any errors they find in the data. Our excellent user support team is good at answering questions, and they pass error information to us right away. Our user base is quite large at this point, and includes people from prominent universities, NGO research units, UN Population and related departments. We do not receive error notices all that often but

it does happen. We correct errors immediately as they are found, and move corrections to the live websites as soon as we can. A Revision History page on the website alerts users to changes or corrections to the data. If the error is egregious, we send a message directly to the users who we know have downloaded that data and let them know what is happening. If, as a data user, you identify an error and let us know about it, we will send you your own IPUMS International mug as a thank-you for helping us ensure the quality of data.

- **How do you decide the coding schemes? Where do the harmonized codes come from and how are those decisions made?**
  We consult international sources for standards on coding wherever we can. Not every variable is complicated enough to have an international standard coding system, but where there is a standardized set of codes, we use those. We also attend meetings and follow discussions on standard recommendations by the UN and other international organizing entities (ILO on employment, UNESCO for education, etc.). We do follow those standards and do the variable organization process in accordance with best practices. Where standards do not exist or where data do not fit the standard, the process is more empirical, so we look at what is common (or common enough) across countries, write thorough documentation, and then harmonize accordingly. With a current collection of nearly 100 countries and more than 350 censuses, we have encountered and incorporated most of the anticipated variation. The project is living and fluid. If a census standard changes drastically, we have to adjust our offering. Two recent examples are changes to disability variable questions as the new Washington Group standards are implemented in a few 2010 round and several 2020 round censuses. The other is a recent ILO change in "status in employment" which alters categories to account for informal work. If changes are significant enough, they sometimes warrant new integrated variables. In those cases, we add documentation to let people know what we have learned about how well (or how poorly) an old standard maps onto a new one for comparative research.

- **What measures for income and for what countries are there in IPUMS?**
  Income is not commonly collected in censuses. In fact, there are just a few censuses that do collect it; the U.S., Brazil, and Mexico are a few. To find those and others as available, the data user can browse the income variable in IPUMS, but it is not available from all the countries. Income is more commonly collected in household surveys and in labor force surveys. We have just been funded to begin incorporating data from international labor force and other household surveys into IPUMS International. We expect to release pilot work on such surveys from Spain and Italy later this year. Our current data from India and Nigeria come from household surveys because we have not yet been able to get approval for the censuses.

- **Do you have full-count census data?**
  For all contemporary censuses, only a sample is accessible through the website. For some historical censuses, full-count data are available. To access the historical censuses, follow the Select Samples page, and then select the Historical tab. Full count data are available for Canada, Denmark, Iceland, Norway, Sweden, United Kingdom, and the United States. Internally, we hold full count data for about 1/3 of the data files available as samples. However, they are in archival storage and not even available to researchers within IPUMS, except for select data checking or methodological work. We are working with some countries to make these higher density and full-count data available via a more secure portal environment, so please keep an eye out for that.

- **How do you check the raw data is accurately collected and recorded when you get it from each country?**
  IPUMS applies a number of data quality checks when processing input data. We check basic characteristics against published total where possible and work hard to identify household breaks accurately by checking characteristics of the household. We empirically verify and document the universe for each variable, label or recode implausible or stray values, and evaluate frequency distributions for accuracy. IPUMS assesses the quality of age reporting in source files using standard indices such as Whipple's Index, Myers Index and the United Nations age/sex accuracy index. We have also undertaken cohort based coherence tests, measuring the consistency of the distribution of an invariant characteristic in two successive censuses such as education.

- **How do we apply survey weights when using different survey years from a single country or multiple countries?**
  Most of the samples that we distribute are systematically constructed by sampling every tenth household across geographically sorted records. Most of the weights, therefore, are just inflation factors rather than differential weights that are seen in DHS or other household surveys. If there are differential weights, they are usually calculated by the national statistics offices rather than buy IPUMS International, and then we distribute the weights that we receive.

- **Can people be linked across censuses? Either directly (via specific ID) or indirectly (via surname/date/place of birth, etc.)**
  At the individual, level, no. We remove all identifying information (names, unique ID's, etc.). Even if there was a unique identification number, it would not be available, and the numbers would not be consistent from year to year. For example, we cannot know if the 10% sample from the 1996 South African census contains the same 10% that make up the sample from the 2001 South African census. IPUMS International data are true cross-sections of the population. That being said, a data user can aggregate using sub-national variables. If the aggregation is done up to a geographical level, longitudinal or panel studies at an aggregate level are possible, but not at an individual level. Our U.S. historical project is constructing linking keys across individuals in the full-count censuses. Since those data are old enough to be identified (names, full addresses, etc.), we have been working with a team of experts to create linking algorithms and to assign linking keys across census years.

- **Wouldn't it be a good idea to also have the global standard definitions such as working population say 15-59, people living below the global poverty line-- also in place?**
  Yes, over the last year we have been talking about what a set of those kind of variables should be. At least a variable defining that, so that the data user could "select by conditional variable", if not versions of employment status that have a couple universe options. Thank you for the confirmation. This is something that we think would be a good idea and hope to make happen very soon.

- **Is there a particular web page where prior webinars are archived?**
  Yes, in addition to the other IPUMS program webinars, this webinar is saved at: https://www.ipums.org/tutorials.shtml. Additional IPUMS International resources are on the on the IPUMS International website

(https://international.ipums.org/international/contact_us.shtml). From there, the user clicks Help at the top of the screen and will be brought to the video tutorials page with You Tube tutorials in a variety of languages, and access to the User Forum.

- **With Canadian public use microdata files and other data now being licensed with an open license, will we see more Canadian data included in IPUMS? https://www.statcan.gc.ca/eng/reference/licence**
  Yes, we are getting better data census samples from Canada, and it is possible we may have time-use and future labor force surveys in future. Thank you to Canada.

- **If an extract I request is still "processing..." and I realize I made a mistake, can I cancel the request? (If not, does the request slow down future requests?)**
  You cannot cancel, but you can revise automatically, and it will not slow down future requests. The revise button is available right away after submitting the extract request. You need not wait for the first extract to finish.

- **How long does it take to add new data release?**
  We typically add 20-25 new census samples every year. The data processing cycle takes almost one year.

- **Can you also tell us how to use fractional weights? When I use them, I get decimal numbers, and I don't know how to tell R that these are different from the standard weights.**
  If we understand your question correctly, you are asking how to handle the situation where your weights are not whole numbers, but rather include decimals. In this situation, some procedures that produce case counts in some statistical packages will round or truncate the weights so that the resulting case counts will always be whole numbers. Often this is because the procedure assumes your weights are frequency weights, which can only be whole numbers. Many widely-used functions for weighted data in R, such as the function svytable in the package survey, do not round or truncate weights, such that they may produce counts that are not whole numbers. Don't worry, this doesn't mean you are doing something wrong! These are still valid estimates of the counts in the population. If you want to present your results as whole numbers, you can use the "round" function; for example `round(svytable(~var1 + var2, mydata))`.

- **Was widowed/divorced inverted for Mexico on that table? Was that how it is coded or just an error in the presentation?**
  The error in the slide has been corrected in the recorded version of the webinar now available at https://www.ipums.org/tutorials.shtml