UN/POP/EGM-CPD49/2015/14

ENGLISH ONLY

UNITED NATIONS EXPERT GROUP MEETING ON STRENGTHENING THE DEMOGRAPHIC EVIDENCE BASE FOR THE POST-2015 DEVELOPMENT AGENDA
Population Division
Department of Economic and Social Affairs
United Nations Secretariat
New York
5-6 October 2015

A NOTE ON DATA CHALLENGES FOR THE DEVELOPMENT AGENDA: OBSERVATIONS FROM IPUMS[1]

Steven Ruggles and Matthew Sobek[*]

[1]**The opinions expressed in this paper are those of the author and do not necessarily reflect those of the United Nations or its Member States. This paper is reproduced as submitted by the author. This paper is being reproduced without formal editing.**

[*] **Minnesota Population Center, University of Minnesota.**

A Note on Data Challenges for the Development Agenda: Observations from IPUMS

Steven Ruggles and Matthew Sobek
Minnesota Population Center, University of Minnesota

The IPUMS projects at the University of Minnesota have been developing data for population research for more than two decades. The hallmarks of that work involve the integration of data collections into consistently coded series and the development of web dissemination systems to deliver those data efficiently to researchers. This integration experience gives the IPUMS team a particular perspective on the immense data challenges posed by the Sustainable Development Goals (SDGs) and the broader development agenda. Researchers will inevitably need to combine data sources to assess goals, to gauge progress, and to address questions that cannot be anticipated now. Measurement of SDGs will require the flexibility of detailed individual-level data to devise consistent measures and develop sophisticated analytical models that explore the dynamics underlying social and economic phenomena. Data access is a necessary—though not by itself sufficient—condition to inform development efforts in the coming decades. Data integration is also required, in order to make maximum use of the world's limited statistical resources.

This note describes our perspective on data dissemination stemming from years of developing systems for use by the research community. This integration work spans both the United States of America and international census and survey data, and more recently has extended to more spatially oriented environmental data. In all cases it involves taking data produced by other organizations for their own purposes and delivering it in an integrated form to third party researchers. As such, there is a somewhat unique position between the data producers and consumers, needing to appreciate the constraints of the one while considering the requirements of the other. The discussion begins by briefly describing the most relevant IPUMS data projects, giving their motivation and some of their key characteristics. A series of observations distilled from our integration work that have particular relevance to the data requirements for the development agenda is also presented.

A. BACKGROUND: IPUMS DATA INFRASTRUCTURE PROJECTS

IPUMS-International. IPUMS-International is the world's most extensive collection of publicly accessible population microdata. The database currently contains information on 614 million people from 277 censuses taken in 82 countries since 1960. Roughly, three quarters of the countries and two thirds of the samples are from the developing world, with seventeen countries on the United Nations list of least developed countries.

Most developing countries outside Western Europe and North America provide good geographic precision, identifying places with as few as 20,000 residents. With a few exceptions, individuals are nested within families and households, and the data contain information about the interrelationships of all members of each residential group. The data also include information on economic activities, ethnicity, educational attainment, fertility, migration and place of former residence, marital status, and consensual unions. Many developing countries provide information about mortality and disabilities, as well as extensive housing characteristics, usually including water supply, sewage, and physical characteristics of the dwelling such as floor and roof materials and number of rooms. For most countries, IPUMS provides microdata from multiple census years. Sample densities are typically 5 to 10 per cent of the national

population, although some are smaller. Because most countries have multiple large samples, it is usually possible to analyze change over time at the sub-national level.

Common variables are coded and labeled consistently and documentation describes comparability issues for each of these harmonized variables. All of this information is presented via a web interface that limits the display to a user's selected samples of interest. A data extraction system allows researchers to select only the variables and samples they require, defining a customized pooled dataset that they download for analysis on their own desktop. The extract system delivers the individual-level data on specific persons, not tables or other summary measures, although there is an on-line tabulator as well. Because variables are harmonized across time and place, IPUMS is optimized to support comparative research.

IPUMS-DHS. The IPUMS-DHS project integrates the Demographic and Health Surveys (DHS) for Africa. The DHS, funded by the United States Agency for International Development (USAID) since the 1980s, is among the most important sources of information on health in the developing world. Although there is considerable consistency in the DHS over time, the data and documentation are strewn across hundreds of files. The sheer logistics of using the data across time and countries poses a significant burden on researchers. Variable coding and question wording changes can easily go unnoticed and be a source of errors.

The IPUMS-DHS project is designed to overcome these limitations and facilitate analysis of DHS data across time and space. The database currently includes data for 76 surveys from 17 African countries and India. The project is expected to be expanded in the coming years to cover the rest of Africa and Asia. Modeled on other IPUMS projects, variables are coded consistently across countries and over time; web-based search and discovery tools display variable availability across surveys; documentation is organized on a cross-survey, variable-specific basis; and researchers can merge files and create customized datasets using a web dissemination system, at no cost. All other attributes aside, IDHS adds great value to the African DHS samples simply by allowing easy browsing of the metadata, to convey the subject content of the various samples.

TerraPop. The newest IPUMS project, TerraPop, aims to integrate three types of data in one system: population microdata from IPUMS; summary data for geographic areas, such as economic indicators or policies; and global raster (grid-cell) data derived from satellite imagery climate models, and other sources. All of these data are spatially referenced, but few researchers have the necessary expertise in the differing data formats to combine them for analysis. Thus, for example, TerraPop allows a microdata researcher to extract a sample of Malawi farmers and append measures of the crops grown in their local area along with indicators of temperature and precipitation. Similarly, an environmental researcher can summarize information on water supply from the census microdata and convert them to raster data to enrich a spatial model.

TerraPop is still under development, but a working prototype that makes several of the necessary transformations and produces data extracts is available. A public system with most of the envisioned functionality should be available by the end of 2015.

B. OBSERVATIONS FROM THE IPUMS EXPERIENCE

Microdata. Population research requires microdata: records comprised of the characteristics of individual persons. Often, microdata are the only way to overcome inconsistencies in summary measures between different places or over time, such as when one country calculates school attendance rates for

different age ranges than another. Published summary data can never be tabulated in all the ways that might conceivably be needed, and they cannot support sophisticated models of individual behavior. To this last point, it is also important that data be organized into households: that the characteristics of each household member are recorded. Individual behaviors are invariably affected by the context of the family and household in which a person lives. In sum, microdata are the most flexible form of data, and researchers need that flexibility to make comparisons and develop causal models.

Metadata. Data are of little use without the metadata to interpret them, yet this essential component is often underdeveloped and it is prone to loss over time. For the IPUMS projects, metadata have typically posed a bigger challenge than the data. Metadata are highly inconsistent across time and between countries, and it can be challenging to acquire the range of metadata components that completely document a dataset. Codebooks are usually fairly straightforward, though they come in various languages and a variety of formats ranging from scanned images to statistical package syntax. But codebooks do not have all the information necessary to fully understand many variables. The most fundamental metadata element is the questionnaire with which the data were collected. It is essential to know what exactly the respondent was asked, especially if one is trying to compare data from different times or places. In our experience, IPUMS have nearly always been able to acquire the questionnaires for censuses and surveys, but have sometimes been unable to uncover accompanying instructions, and often have little information on how missing or inconsistent data were edited. Sample design information can also be difficult to obtain and hard to interpret. IPUMS translates all questionnaire text into English for our metadata system, but also provide access to the original-language documents to retain potential nuances of meaning.

Data Integration. Given the scope of the development agenda, researchers will inevitably need to combine data from different sources over time and across countries. For two datasets collected using the same instrument and processed similarly, integration can be fairly easy. But where that simple scenario does not obtain, the costs can be high in terms of effort and results can be error-prone and difficult to replicate. For core data collections that will be used by many researchers, the investment in formal data integration is warranted.

The core activity of integration is harmonization of codes at the variable level, but such harmonization risks creating a lowest common denominator dictated by the least detailed dataset in the collection. As always, it is not possible to fully predict what will be important to future analysts. To prevent loss of information, a composite coding system is used where the first digit of the variable is fully comparable but where trailing digits retains category detail present in a subset of samples. Access to the original non-integrated variables is also provided.

While harmonization of codes is important, data integration goes much beyond this. Intelligent interpretation of the data requires integrating the metadata as well: the labels, questionnaires, variable descriptions, and sample design information. Comparability issues need to be made readily apparent to researchers. It is impossible to anticipate all the uses to which the data could be put, so the goal should be to highlight major issues while providing the tools for researchers to explore the metadata more deeply—including the original documentation of the input data prior to integration.

Dissemination. Web-based data-access tools are essential. These tools must provide users with the capacity to design customized datasets that pool data across time and space to simplify analysis of change and cross-national comparison. Perhaps the most valuable contribution of an integrated web-based dissemination system lies in data discovery. Displaying variable availability across the data collection efficiently conveys to researchers what kinds of questions can be addressed in particular times or places. That availability list should provide ready access to the deeper variable-level documentation, so researchers can quickly ascertain whether their key variables were derived from question wording that supports their needs. The data extraction must be seamlessly integrated with metadata browsing functions.

The logistics of file management and data manipulation are sand in the gears of research and should be minimized. A system that provides pooled datasets combining data across times and places in the format researchers want is more preferable. It is counterproductive to provide dozens or hundreds of variables that are not needed but which must be waded through to conduct any given analysis. Customized data extracts containing only the necessary variables and subpopulations are much more efficient.

Historical depth. Time is an essential dimension to incorporate into any data strategy. Without knowing how things stood in the past it is impossible to ascertain whether a situation is improving or worsening. The greater the historical depth, the better our understanding will be of the current trajectory and the greater the scope for learning from past natural experiments. The necessity of some degree of historical depth implies the need for data integration of current and future sources with data collected in the past. New surveys oriented to the SDGs will need to be bridged to older sources using common variables, proxies, and the like. Comparability to those past data collections needs to be a key consideration in the design of new instruments.

Censuses. Census data are the backbone of population research and it is critical that researchers have access to high-density microdata samples from developing countries. Censuses provide the sampling frame for surveys and are the only source that covers the entire population, including members of collective dwellings. Only censuses have sufficient cases to study smaller subpopulations, such as particular ethnic groups or the very old. Census data are therefore needed to ascertain how broadly selected development goals are achieved across all segments of the national population. The size of the census samples also enables robust measurements at the subnational level as well as over time, since most countries have census microdata going back at least two or three decades. Although they have limited subject content compared to purpose-designed surveys, censuses can provide contextual details for survey analyses, and they can provide measures that let researchers extrapolate survey results to the broader population. For privacy reasons low-level geographic detail must be suppressed in public use census microdata samples, but we hope some countries will agree to a restricted-use data enclave arrangement where careful vetting of output will enable access to full-count data with complete geographic information.

Geographic integration. Geographic integration is generally necessary to support analyses over time at the subnational level. Not only may some regions lag behind other; national borders are not always the most relevant geographic units for phenomena related to culture or environment. Over decades-long timespans, administrative boundaries within countries often change, and surveys may impose their own geographies at which their designs yield representative results. The practical solution is to construct integrated geographic units within which boundary changes occur, but which are themselves stable over time. These units are necessarily larger than those in the original sources, so it is necessary to provide the sample-specific geographies as well to support more focused research. Devising such integrated units is labour-intensive and, critically, depends on acquisition of GIS boundary files or creating them from maps. Considerable efforts have been invested in creating these units at the first and second administrative levels for the countries and times where census microdata are available, but the starting point is the modern boundary files produced by the national statistical offices. These boundaries, should be offered as a public good by all countries, as are the boundary files constructed for IPUMS and TerraPop.

C. CONCLUSIONS

Data integration reduces costs over the long run by minimizing the costs of data analysis. Because high-quality data integration is rare, most of the world's data are expensive to use for international

comparisons or analysis of change. Investment in data integration is not cheap, but the costs are small compared with the costs of data collection. Moreover, integrated data and metadata reduce the potential for error and increase the replicability of results. Accordingly, it is felt that all the major classes of population data and administrative data should be integrated across countries. In general, highly detailed disaggregated datasets—preferably at the individual level—are much easier to integrate than are aggregated data, and they are much more flexible for analysis.

International standards for data collection and questionnaire format are invaluable, and greatly improve the prospects for data integration. Metadata are critical, and they should not be an afterthought. All data producers should produce comprehensive machine-processable metadata, including full descriptions of data collection instruments and processes. If the format of these metadata were standardized across countries, it would greatly reduce the costs of data integration. This would not, however, substitute for a systematic programme of data integration. As long as data are collected and processed by different agencies, incompatibilities will remain that must be sorted out.