21

Using and Improving International Microdata for Historical Research

Lisa Y. Dillon
Gunnar Thorvaldsen

Vast amounts of microdata from many countries are already available for historical population research. Plans to digitize more source material and rescue existing data files suggest that the number of records at our disposal will increase significantly during the next decade. To truly maximize the research potential of these powerful new databases, they must not only be made available, they must also be made compatible—across time and across national borders.

Data Harmonizing Projects

Several research teams around the world have taken up this challenge of harmonizing international datasets. Under the auspices of the regional economic commissions of the United Nations, several large projects have collected census data for a particular region of the world. CELADE (the Latin American Demographic Center) and PAU (Europe's Population Activities Unit) are two examples. In 1999, the Minnesota Population Center launched an international database project, bringing the technical expertise acquired in harmonizing 150 years of U.S. census data into

Dr. Lisa Y. Dillon is Researcher and Research Coordinator, Institute of Canadian Studies, at the University of Ottawa. Dr. Gunnar Thorvaldsen is professor of history and manager of research at the Norwegian Historical Data Centre, at the University of Tromsø, Norway.

the international arena. The IPUMS-International project has begun to inventory the world's known microdata and, where access and confidentiality rules permit, they will begin to harmonize both historical and contemporary census data using the United Nations guidelines on data compatibility as a starting point.[1]

In addition, the IPUMS-International project has also started to document the contents of census variables around the world in greater detail than what can be found in the Goyer handbooks.[2] This documentation will be the basis for the huge task of standardizing censuses multilaterally in the same manner as has been done with the U.S. censuses.

One of the strengths of the IPUMS-USA database is that it was designed for research. Although built by historical demographers, from the very beginning Steven Ruggles and the team at the Historical Census Project solicited advice from scholars on how to structure the IPUMS data to be useful for research in other disciplines, as well. Test versions of the database were made available and researchers were encouraged to use them and report any analytical problems they encountered.

The common denominator of the IPUMS-International variable coding system will be much broader than the the existing IPUMS formats, however, since in many countries the censuses contains variables—such as religion and ethnic categories—which are not found in the U.S. census.

While the CELADE, PAU and IPUMS-International work to harmonize multiple variables across country boundaries, another team of scholars headed by Marco van Leeuwen and Ineke Maas has been working to design a historically and geographically compatible set of occupation codes. Although UNESCO (United Nations Educational, Scientific and Cultural Organization) issued occupational coding guidelines in 1958, 1968 and 1988, these coding schemes do not work well for the imprecise, historical titles found in earlier sources. Building on the ISCO68 standard from the International Labour Office (International Standard Classification of Occupations, Geneva 1968), the Historical International Standard

---

[1] For more information on the CELADE, PAU and IPUMS-International projects see Chapters 17, 18 and 20.

[2] Doreen S. Goyer, Eliane Domscheke and Gera E. Graaijer, eds. (1983). *The Handbook of National Population Censuses*, in 4 Vols., Westport, Connecticut: Greenwood Press.

Classification of Occupations (HISCO) standard incorporates codes for a number of obsolete titles, a status code, temporary code, a product code and a number of rules to make the coding procedure more uniform across time and space.

A preliminary version of HISCO is now available and, as with the IPUMS, van Leeuwen and Maas have been cooperating with other researchers, encouraging test usage and advice on revisions.

Comparative Research Using Harmonized Microdata

As these international data harmonizing projects move forward, they need the help of social science historians to construct variable formats that are flexible enough to accommodate the question and response variations that occur in all but the most basic variables.

Each of the authors of this essay has used microdata from their own country—Canada and Norway, respectively—and merged it with IPUMS microdata from the U.S. They also tested the HISCO occupation codes in both datasets. What follows is their report of some of the problems encountered in handling and analyzing the merged datasets.

Harmonizing Norwegian and U.S. Microdata

As a preliminary test of the difficulties of merging data from the Norwegian and American censuses of 1900, Thorvaldsen used the municipality of Tromsøsund, which contains some 5,000 inhabitants in a mixed rural and urban setting. The aim was to use IPUMS variables where compatibility with these could be obtained, while constructing new ones where necessary. This meant that some variables specific to Norway, such as religion had to be kept, while adding others in order to accommodate the encoding of occupations according to the HISCO standard. Also, it was aimed to use the software developed for the IPUMS by Steven Ruggles to include a number of constructed variables, most notably those specifying family and household relationships.

Technically, the job was done with MS Access, using a number of auxiliary tables to convert data values via SQL queries.

For some variables the task was straightforward. But IPUMS uses different digits than the Norwegian standard to encode marital status, and this was converted via a simple look-up table.

Harmonizing the birthplace variable was especially complicated because the level of detail is greater in the Norwegian than in the US censuses. While the latter only contain information on the country of birth for immigrants and US state for the native born, the Norwegian birthplaces are specified as municipalities. Therefore, while it makes sense to use the IPUMS national birthplace codes for immigrants to Norway, an additional variable with the domestic municipality codes must be added for the native-born Norwegians.

For occupations, the system built on the contemporary coding of the late nineteenth century censuses will be kept. These encode main and secondary occupations according to trade and position in the social hierarchy. Additionally, the historical version of the HISCO codes occupations in a more detailed way, especially stressing the function of each specific occupation. While some of the original Norwegian codes could be used straightforwardly as a basis for HISCO codes, most titles had to be dealt with independently, making a judgement about each of them.

The Norwegian system for coding family and household positions classifies the relationship of each individual to the head of household in ten groups. A new encoding of the family and household status variable had to be created. The constructed IPUMS codes specified in the family interrelationship variables, however, require that each individual household be encoded separately. One final problem is that Norwegian name customs are different from the English, a patronymic based on the father's first name indicating relationships rather than using surname similarity. In spite of this, the location of spouses, mothers and fathers could be encoded automatically in most households.

Harmonizing Canadian and U.S. Microdata

Drawing upon approaches developed for the IPUMS project, Dillon first integrated the 1850 and 1880 U.S. census microdata with a similar set of data for Canada in 1871. She later added the 1871 Canadian sample, 1901 national Canadian Census Sample, 1900 U.S. public use sample and a preliminary subsample of the 1870 U.S. public use sample. Her purpose was to explore whether the

Canadian and U.S. political boundaries were paralleled by national distinctions in the living arrangements of the elderly and their children.

The resulting 1870/1 - 1900/1 data series includes over 600,000 individuals.[3] The census samples from the two countries offer information on a range of similar variables, including sex, marital status, age, relationship to household head, place of birth, occupation, place of residence, and, in some of the samples, ethnicity, religion, race, property-ownership, relationship to the means of production, and income.

At first glance, integrating the four microdata sets seems relatively straightforward, for the samples bear strong similarities in enumeration practice, sampling design, data collection and data organization. In many instances, Dillon used IPUMS codes, making appropriate changes to the IPUMS relationship-to-household head and birthplace codes. Two challenges faced in harmonizing the data owed to sampling differences and occupational classifications in the currently available version of the 1900 U.S. sample. These will be alleviated with the completion of the new 1900 U.S. sample, now being constructed by the Minnesota Historical Census Project, which will collect dwelling-level information and will include both alphabetic occupation strings and detailed level occupation codes.

Dillon encountered few difficulties harmonizing the 1901 Canadian census data with corresponding datasets as the Canadian Families Project (CFP). In fact, the CFP has made use of some IPUMS codes from the beginning. The chief challenge in integrating the 1901 and 1871 Canadian census microdata with their U.S. counterparts has been in reconciling information pertaining to household relationships. Since the 1870 U.S. census did not ask relationship-to-household head, the IPUMS uses an imputation procedure to create a relationship-to-household-head variable for this PUMS. Steven Ruggles is currently working with the Canadian Families Project to create a similar variable for the 1871 Canadian PUMS, in this case borrowing relationship values from

[3] The 1870 and 1900 U.S. Public Use Samples (PUMS) are available from the Minnesota Historical Census Projects; the 1871 PUMS was created by Gordon Darroch and Michael Ornstein at the Institute of Social Research, York University; the 1901 census data set, which will become publicly available in 2001, was created by the Canadian Families Project.

the 1901 Canadian census microdata and adding in information on religion and ethnicity. The researchers will then be able to analyze whether using ethnicity and religion (which are not available in the U.S. PUMS) as matching criteria makes a significant difference in imputing household relationships.

Further challenges in working with integrated Canadian and U.S. historical census microdata arise at the stage of historical analysis. For example, geographic disparity is highlighted in the the 1871 Census of Canada which enumerated the four provinces—Ontario, Quebec, New Brunswick and Nova Scotia—whereas the 1870 U.S. enumeration includes Americans from coast to coast.[4] This geographic disparity forces the researcher to question how to structure comparisons between the two countries in 1871: should we compare Canada as it was then constituted to the whole of the United States or only to the eastern states? By 1901, the Canadian census included persons from Victoria to Halifax. Nevertheless, the population of western Canada in 1901 was much smaller than the population of the western United States; the flood of immigrants, among them Americans, to the Canadian West was yet to come. While the count of cases in western Canada is large enough to permit general comparisons with the western U.S., analysis of smaller subgroups such as the elderly becomes very difficult. The Canadian Families Project has attempted to redress this problem by entering 100% samples of selected western communities.[5]

Canada, the U.K. and the U.S.

The near future holds rich possibilities to develop more integrated Canadian-U.S. census microdata sets. Dillon and her colleagues with the Canadian Families Project and le Centre Interuniversitaires des Etudes Québecoises (CIEQ) have established

---

[4] Manitoba and British Columbia were enumerated in 1870, but these censuses have not been sampled in national microdata files.

[5] Dillon has detailed the integration of the Canadian and U.S. historical census microdata in two publications, "Integrating Canadian and U.S. historical census microdata: 1871 and 1901 Canada, and 1870 and 1900 United States," *Historical Methods* (forthcoming) and "Integrating Nineteenth-Century Canadian and American Census Data Sets," *Computers and the Humanities*, 30 (1997): 381-92.

a partnership with the Church of Jesus Christ of Latter-Day Saints (LDS) to check and clean the LDS 100% microdata set of the 1881 Canadian census. In return for this work, Canadian scholars will retain a copy of the 1881 microdata for historical research.

The Canadian partnership with the LDS is paralleled by a similar arrangement between the LDS and the Minnesota Historical Census Project. The University of Essex History of Work project has also obtained and enhanced a 100% sample of the 1881 census of England, Wales and Scotland. The analytic usefulness of these 100% samples would be increased exponentially by integrating the three. In addition, the Institute of Canadian Studies is launching a project to create samples of the 1911, 1921, 1931, 1941 and 1951 censuses of Canada, to complement existing PUMS of the 1971 to 1991 Canadian census and an electronic file of the 1961 Canadian census. The plans for this project include a phase in which the twentieth-century Canadian census samples are integrated with their counterparts in the United States. The accompanying chart displays the large number of variables held in common by the late nineteenth- and twentieth-century U.S. and Canadian census samples. Standardizing PUMS from the two countries will, for example, facilitate research on the impact of Canada-U.S. trade on North American families. Much research has been devoted to Canadian-U.S. political and economic relationships, but until now, scholars have not been able to study how fluctuations in Canada-U.S. cross-border trading and migration policies may have affected their populations. These integrated microdata would also provide researchers with the opportunity to study Canadian-U.S. similarities and differences by examining cross-border communities from the Pacific Northwest coast to Maine and the Maritimes. Integrating both the nineteenth- and twentieth-century Canadian microdata samples with the U.S. IPUMS allows scholars in both countries to broaden their analytic framework considerably.

Future Directions

Through international co-operation, researchers can design the comparable and integrated census microdata needed for social science research. Organizations such as IMAG provide valuable forums for researchers to discuss similarities and differences in their approaches to and use of these microdata.

Like the routinely-generated historical documents upon which they are based, census microdata are constructed bodies of information and inevitably reflect the priorities and assumptions of their creators. Only through international dialogue can database creators and users begin to understand and consider modifying these priorities and assumptions. The long-term historical and trans-national research based on this international research data infrastructure will provide a crucial baseline for understanding the dynamics of poverty, race relations, sex discrimination, migration and aging— all issues which continue to affect world populations.