1 Introduction

International Historical Microdata: A New Resource for Research and Planning

Patricia Kelly Hall
Robert McCaa
Gunnar Thorvaldsen

Like the trolls of fairy tales, historians hunt everything human—everyone, everywhere if global history is our ultimate ideal. When considering whether global history is feasible, one crucial requirement must be to assemble at least the most basic information about all members of human society. Such information can be found in nominative lists, such as the census, which report information on the sex, age, marital and family status, occupation and birthplace of each individual. This type of source material can be found in most countries, and usually the whole population is included.

This book describes the availability of such basic population source material for many countries, how it is transformed into data registers and how these data can be used to write the history of social groups from the family level, through regional and national analysis to international, comparative surveys.

Patricia Kelly Hall is Coordinator of the Integrated Public Use Microdata Series (IPUMS) project at the Minnesota Population Center (MPC). Dr. Robert McCaa is Professor of History at the University of Minnesota and a principal investigator of the IPUMS-International project at the MPC. Dr. Gunnar Thorvaldsen is professor of history and manager of research at the Norwegian Historical Data Centre, at the University of Tromsø, Norway.

What are Microdata?

Microdata contain all the information for each individual included in a population register, survey, census or other data collection instrument. Each person and household is a separate record usually composed of numerically coded data. Although straightforward counts or simple analyses can be done with a computer spreadsheet, most samples require analysis with a statistical program.

The majority of the datasets described here are drawn from a country's national census and they provide the best examples of how microdata differ from what has been available to researchers in the past. Information from the census, like the source material for all the microdatasets in this volume, is not new. Printed reports of a census summarizing characteristics of the population have regularly been made available within a few years after the census was gathered. Printed census volumes contain tables—summary information that answers one specific question, such as "how many people of each age group lived in place X." They describe only one or two characteristics of each person enumerated.

Microdata files usually are samples of 1 out of every 100 persons drawn from the complete enumeration. They provide all the information collected for each person in the sample. Instead of a table giving a count of how many people in a given locality in a given year shared one characteristic, census microdata give *all* the responses of individual persons and households transcribed from the census manuscripts. A census microdata sample thus allows researchers to know, simultaneously, all the personal characteristics of every individual in the sample.

This is the power of microdata and the characteristic that is shared by all the datasets described in this book: they provide information on all the known characteristics of an individual. Many of these datasets link household information with individual information, thus allowing researchers to know the characteristics of the indivdiual and of everyone with whom they reside.

Accessing and Analyzing Microdata

The mainframe computer is no longer a necessity for most social science research. The revolution in computer technology has given individual researchers access to ever more computing power on their desktops, making the file size of large data sets a non-issue. For many users, a personal computer and internet access are the only resources necessary to carry out research with any of the datasets described here.

Access to the data was the first barrier to collapse. With an internet connection and a few simple file transfer protocol (ftp) commands, researchers can now download data files from anywhere in the world. The recent introduction of the World Wide Web has made the file transfer process for many databases even simpler, completely eliminating the user's need to know ftp. Some of the larger databases, such as the Integrated Public Use Microdata Series (IPUMS) in the United States, have even initiated web-based extraction systems that allow users to select the data they want simply by checking boxes on the screen.

Proficiency with a statistical package was another obstacle preventing scholars from using microdata as a primary source. Here, again, technological advances are improving access. Programs such as SPSS, SAS and Stata have now been adapted for use with Windows-based operating systems. This means that the power of these packages can now be deployed easily by clicking a mouse to select needed procedures or data transformations – a much easier process than understanding the underlying syntax of hundreds of command statements.

Overview of the Datasets

The essays presented in this handbook cover a variety of national and international datasets—both historical and contemporary—from different time periods and different parts of the globe.

Historical Datasets

Argentines pioneered the construction of nationally representative machine-readable census microdata samples for nineteenth century populations. Completed in 1967 with a total

of 200,000 individuals for the censuses of 1869 and 1895, the Argentine samples remain a model for later efforts due to their thorough documentation, integrated coding schemes, and systematic procedures.

The Canadian Families Project's 1901 Census sample offers a close look at labor force participation and changing language patterns in a multi-ethnic population nearly unrivalled by that of any other nation.

In their innovative drive to collect and disseminate microdata, the Danish Data Archive continues to draw on the resources and time of amateur historians throughout Denmark. Their collaboration suggests how historical microdata projects might draw on a number of constituencies for collegial support and data processing.

Besides digitizing and making available important census microdata, the 1851 and 1881 Censuses of Great Britain pioneered non-traditional partnerships (namely, with the Church of Jesus Christ of Latter-Day Saints) that rescued records and deepened an already extensive database while benefiting both parties as well as scholars across the world.

The Historical Sample of the Netherlands contains individual data culled from birth, marriage and death certificates as well as population registers. Compiling life history data in this manner allows the project to draw on and collate a rich mixture of historical microdata documents.

At the University of Tromsø, the Norwegian Historical Data Centre is currently digitizing a collection of population microdata that extends back into the seventeenth century. The Centre has already made a great deal of nineteenth-century census data available for researchers.

The Stockholm Historical Database in Sweden has been digitizing population records since the 1970s. Capturing demographic change in the city from 1878-1926, their database provides an especially in-depth look at people through microdata. And at Umeå University, also in Sweden, an extensive set of church registers--digitized by the Demographic Data Base—chronicles social change across the eighteenth and nineteenth centuries across a large area.

The extraordinary database for the Granduchy of Tuscany is one of the most comprehensive currently available to

historians. The database refers to a population of 1.5 million individuals constructed from a systematic name-by-name mining of census lists, parish registers, land cadasters, and even child abandonment records.

The Integrated Public Use Microdata Series (IPUMS) includes samples of all the available censuses of the United States from 1850 through 1990. The database was constructed by the Historical Census Project at the Minnesota Population Center. The IPUMS data are available free to researchers, who can request and receive data extracts and documentation over the World Wide Web.

Contemporary datasets:

China is an excellent example of a country with a great wealth of highly comparable census microdata of recent origin (1982, 1990) and restricted access. With the integration of the 2000 census sample, an enormously valuable database of as many as thirty million people is likely to become available to researchers.

Colombia, the third largest country in Latin America, has the longest string of complete national census microdata currently available. Begun in 1964 with a two-percent sample of individuals, the series comprises the complete datatapes for censuses in subsequent decades. Moreover the Colombian statistical authority, DANE, is eager to construct new high-density samples for inclusion in the international census integration project.

Computer-readable microdata from France provide an opportunity for extensive comparative work. In their data, detailed family structures and complex defintions of occupations might prove especially useful for transnational research.

Korean census data, archived at the East-West Center at the University of Hawaii, track the nation's last forty years of growth. The Center's work demonstrates the importance of comparative historical microdata scholarship—the Republic of Korea data base allows scholars access to the burgeoning effects of one nation in Asia's integration into the world economy.

The Mexican Statistical Bureau (INEGI) prepared its first machine-readable census microdata sample for researchers for

the 1960 census. In the 1990s, commercialization made Mexican census microdata samples even more widely available. In 2000, with almost 10% of the Mexican-born population resident in the United States, integration of census microdata will provide researchers a valuable opportunity to conduct cross-national investigations without the hurdle of incompatible metadata.

The Latin American Demographic Center (CELADE), sponsored by the United Nations, was the first institution to undertake a large-scale census integration project. During its lifespan from 1959 to the early 1980s, CELADE's OMUECE project had amassed a collection of 61 national samples on some two dozen Latin American countries, dating from the 1960 round of censuses.

The European Commission for Europe's Population Activities Unit has developed a path-breaking harmonization of census microdata for almost two dozen European countries. While sample designs vary from country-to-country and the database currently lacks chronological depth, it remains one of the most extensive census microdata projects currently underway.

The two final essays move beyond describing individual datasets to ways in which these new resources can be used in comparative research.

The Great Britain Historical GIS Project's work suggests that historians of census microdata need to think carefully about the delimination of census spaces—an often unquestioned foundation of census data collection and histories. Growing alongside the research on the 1881 British census, the Great Britain Historical Geographic Information System offers scholars a new look at old data. By providing definitions of boundaries from the past, historians of the 1881 British census as well as others can now take on spatially-referenced census data in new ways.

Lisa Y. Dillon and Gunnar Thorvaldsen are two scholars who have used national microdata bases from two countries to do comparative analysis in their own research. Although a few problems do occur, they show from their own experiences that these problems are surmountable. Dillon and Thorvaldsen suggest that it is precisely this kind of research testing by scholars in many disciplines—and the sharing of these

experiences with the project teams building the databases—that will yield more powerful and flexible international microdata bases in the future.

The final section of the *Handbook* presents more detailed specifications of the datasets described in the essays as well as an inventory of the world's known microdata. The inventory is a work in progress, part of the IPUMS-International effort to catalogue the world's known and available microdata.

Historical microdata provide social scientists with a powerful and flexible resource. Researchers can custom-design their own tabulations that do not depend on the categories pre-selected for them by the Census Bureau in the past. Even the modern Bureau is incapable of predicting or presenting all the tabulations required by researchers. The microdata often make it possible to overcome inconsistencies in published time series where, for example, the minimum age or sex of persons subject to a tabulation may have changed. The microdata also enable more sophisticated multivariate analyses exploring the relationships between characteristics at the individual or household level.

There is a large potential audience of researchers wanting simple background statistics for largely non-quantitative applications: historians seeking context for local studies, teachers desiring descriptive numbers for lectures and discussion, and journalists seeking summary data addressing their shifting concerns.

To those approaching the use of historical microdata for the first time, our aim is twofold: first, to explain the revolution in ease of access and analysis that has occurred in this field in the last twenty years and, second, to provide an introduction to the datasets that are either currently available for research purposes or are in production. Internet addresses and bibliographies are included to get people started.

To those who are already using microdata for their research, our goal is to encourage more comparative analysis — either across time or across political boundaries —by exploring more of the rich data resources described here.