ESA/STAT/AC.84/26
1 August 2001

English only

Symposium on Global Review of 2000 Round of Population and Housing Censuses: Mid-Decade

Assessment and Future Prospects

Statistics Division
Department of Economic and Social Affairs
United Nations Secretariat
New York, 7-10 August 2001

Archiving Census Documentation and Microdata: Preserving Memory, Increasing Stakeholders

Wendy L. Thomas and Robert McCaa

Session 4
Maintaining census related activities during intercensal years

UNITED NATIONS SYMPOSIUM
GLOBAL REVIEW OF 2000 ROUND OF POPULATION AND HOUSING CENSUSES:

MID-DECADE ASSESSMENT AND FUTURE PROSPECTS

August 7-10, 2001 / UNSD-NY.

Archiving Census Documentation and Microdata: Preserving Memory, Increasing Stakeholders

Wendy L. Thomas and Robert McCaa
University of Minnesota Population Center
wlt@pop.umn.edu; rmccaa@umn.edu

1.0 Introduction

The preservation of various types of census materials must be raised early in the cycle of census activities. Well-preserved data and documentation contribute to effective data collection, dissemination, planning, and future use of the population census. The ability to learn from past processes, identify strategies that contribute to a successful census, retain and build on core activities and structures from previous censuses and effectively apply census data to current and future issues is all dependent upon the preservation of census data and the materials related to the collection and processing of that data.

In an ideal world with unlimited resources the questions of what to preserve and how to preserve it would be easier to address. Unfortunately this is not the case and even in the wealthiest of countries the cost of preservation, and questions surrounding the means of preservation, have a profound impact on what materials are preserved and in what format. The purpose of this paper is look at the types of data and documentation accumulated during the census process and explore the benefits of preserving these types of documents in terms of informing future censuses and data users, ensuring appropriate preservation formats, and identifying stakeholders who may be an effective force in lobbying for the preservation of various classes of documents.

Classifying materials for preservation in terms of their future impact and anticipated use is useful for identifying the trade-offs in preservation decisions for individual countries. By coupling this type of materials lists with an inventory of the available technology, personnel and knowledge within a country to process materials for preservation, governments will have the information necessary to enable them to make informed preservation decisions. The use of a questionnaire to elicit information on the available infrastructure for preservation within a country may also bring to light options for cooperative services or a profile of appropriate technologies for a variety of situations. The ability to not only determine what will be preserved, but also what will not be preserved, based on an understanding of the long-term impact of the information contained in the document is instrumental in developing a long-term census preservation policy that will meet the needs of future generations.

2.0 Long-term preservation of data and documentation

2.1 Definition of long-term preservation

Long-term preservation takes on a new meaning with electronic records. “Archiving” is a term used both by computer/information technology specialist and archivist, yet conveys different meanings to these two groups. “Archiving” in the world of computing refers to inactive or off-line storage. To archivists “archiving” means to preserve an information record in a format that is independent of its production environment and to protect that record from loss, alteration or deterioration.

For archivists, well-preserved electronic records have the following characteristics. (Dollar, 47-57) They are:

- **Readable**. In short, they are undamaged and the bit-stream can be processed either the machine that created it, the machine that is storing it, or the machine on which it will be stored.
- **Intelligible**, having sufficient metadata to interpret the 1s and 0s of the bitmap image. In other words information regarding the compression algorithm and the byte order. This is similar to the file extension TXT denoting a 7-bit ASCII text file. Without this basic level of metadata the record is for practical purposes unintelligible.
- **Identifiable** in that a unique ID or attribute can locate them.
- **Encapsulated** so that all the information associated with a record (its metadata and linkages) exist as a single logical or physical entity
- **Understandable** through the provision of full metadata.
- **Reconstructable** in terms of the logical, physical and intellectual content.
- **Authentic** records. “Archival science defines authentic records as being what they purport to be – reliable records that over time have not been altered, changed, or otherwise corrupted.” (Dollar, 54)

It is important to keep this concept of preservation in mind when assessing the value of preserving particular census records and in determining the costs of distribution, storage and long-term preservation.

2.2 The value of preservation

Much has been written on the importance of organizing and coordinating the process of census taking within and between countries (United Nations, Department of Economic and Social Affairs, Statistical Division. *Handbook on Census Management,* 2000). Numerous intergovernmental and non-governmental agencies provide support and assistance for this process. Emphasis has been placed on planning, data collection, methodologies, product preparation and dissemination. The value of a strong archival program lies not only in preserving the actual data, metadata and data products for future use, but also in its ability to contribute to future census and statistical activities.

Given the periodic nature of census taking, maintaining records on how specific activities were performed can inform future census processes within a country, allowing agencies

to learn from past processes and strategies. This is particularly important in countries that do not have and cannot afford to have a permanent office for the census. Carefully selected and preserved records can provide detailed information on the planning process, specifications of collection, and insight into why certain decisions were made and how effective particular activities were. In particular, it is these types of country specific processes and approaches that can assist in retaining and building on successful core activities and structures.

Preservation and communication of information on data quality and process evaluation is of value for informing future census activities and is essential for the informed use of census data. Communicating information on the reliability, limitations and strengths of the final data allows users to understand the impact of any procedural changes on any analysis they may wish to perform. This is the type of information that should be encapsulated through logical or physical links between the census data and the procedural metadata in the preservation process.

2.3 Costs of Preservation

The cost of preservation is an issue for all countries. Recent discussions of retention schedules for the 2000 U.S. Census elicited numerous responses from various stakeholder groups concerning both the preservation of original forms and intermediary process output. The cost of preserving original enumeration forms in various formats and the associated cost of making these identifiable for future users was one of the key factors in negotiating a final retention schedule.

In countries without permanent census offices and/or permanent national archives structures, the cost of preservation becomes a major issue. By looking at these costs early and including them in the discussion of the overall costs in undertaking a census, additional options for allocating funds may be found. For example, the way in which census data is captured and prepared for dissemination can reduce the cost of creating a preservation quality record. In addition, capturing and retaining procedural information as it is produced and creating the logical or physical links to emerging data collections, increases the likelihood of preservation while reducing the cost of reconstructing valuable metadata information.

Early discussions of the costs and future value of information preservation allow for both informed decisions and the opportunity to discuss cooperative long-term preservation possibilities in a timely fashion.

3.0 Determining What to Preserve

3.1 Preserving the products

The essential elements of any census in terms of preservation are the resulting data and basic documentation. How that data is identified and defined varies by country. Issues of confidentiality and security play a major role in determining not only who should have access to the microdata and enumeration forms, but also whether that information should

be retained at all. Increasing the availability of microdata contributes to the likelihood that these data will be preserved.

Access to microdata is being made available by an increasing number of countries in a variety of forms: Public samples, scientific samples (restricted to a few carefully screened projects), and through data enclaves where the user works in a secure site and output is tightly controlled. From 1985 through 1994, of 153 countries with populations of one million or more, 134 conducted enumerations in the 1990 round of censuses. 94% of the world's population was counted. 54 countries provided researchers access to anonymized census samples of individuals and households. Some countries restricted access to a single investigator or research facility, but what is remarkable about the 1990s is not only the globalization of the census, but the growing acceptance of anonymized samples as statistical instruments. These trends are continuing in the 2000 round of censuses (1995-2004).

The approach used in the United States of providing public samples of sizes ranging from 1-15% for various area types supports a wide range of research at both the local and national level. In addition, the release of the data from restricted status after 73 years has resulted in a number of projects to make this data accessible to the public in digital format. The most noted of these is the Integrated Public Use Microsample (IPUMS) project. This project, begun in 1992 at the University of Minnesota, integrates sixty-five million microdata records for the United States. Conceived by Steven Ruggles, founding director of the Minnesota Population Center, and funded by the National Science Foundation and the National Institutes of Health, IPUMS integrates the decennial censuses of the United States, dating from 1850 to 1990. The first version of the IPUMS database was released on tape in 1993 and by 1995 via the Internet. Thanks to the expansion of the Internet, the data distribution problem was easily solved by means of a web site driven data dissemination engine (http://www.ipums.org). The IPUMS database, distributed free of charge via the Internet, quickly established itself as one of the three most frequently cited data sources in population research about the United States.

In October 1999, with major funding secured from the National Science Foundation, a global effort was inaugurated, dubbed, IPUMS-International. With the cooperation of national teams of investigators, the IPUMS-International consortium proposes to integrate census microdata for more than a dozen additional countries, with at least one from each continent. Historical census microdata for Canada, Norway, Great Britain, Argentina, and Costa Rica will be included in the database as well as those for the United States. Contemporary microdata for Colombia and the United States will be integrated along with those for France, Brazil, Mexico, Vietnam, Kenya, Great Britain, Hungary, Spain, and others. Based on a prototype developed with the cooperation of the Colombian National Statistical Office (Departamento Administrativo Nacional de Estadística, or DANE), country teams of experienced census data-users are being formed to advise on how to harmonize the national census concepts using international norms.

The creation of public use samples is being used by a variety of countries to increase access to microdata. Software such as the Integrated Microcomputer Processing System (IMPS and its successor CSPro), a system for data processing of censuses and surveys, developed by the International Statistical Programs Center of the U.S. Bureau of the Census, facilitates the dissemination of microdata samples by providing tools for cross tabulation, electronic map production and other basic analysis, thereby reducing the cost of producing these products for individual countries.

Examples of distributed microdata public use samples include Vietnam, which has released a 3% sample of the 1999 Population and Housing Census, with the intention of producing a full 100% sample at a later date. Mexico has released a 10% sample designed to yield valuable information at the level of municipalities of 100,000 or more in size. France has released 5% samples for 1962-1990. Likewise the Central Bureau of Statistics of Kenya has prepared a mega-sample of the 1999 enumeration (with a maximum density of twenty percent) to complete its impressive series of samples for 1969, 1979, and 1989.

These collections not only provide data in a preservable format, they include a range of metadata. The documentation is extraordinarily complete, and includes details on every aspect of the census from earliest preparations to the final publication of tables. The discussion of sampling is particularly noteworthy.

A growing list of countries is offering data in the REDATAM format (developed by the United Nations Demographic Center for Latin America and the Caribbean, CELADE), as a way of storing microdata and making them useful to researchers and administrators who need small area statistics.

REDATAM "**RE**trieval of **DAT**a for small **A**reas by **M**icrocomputer" was originally conceived of as a low cost data retrieval computer program and has grown into a concept that involves a proprietary database format, as well as a software development system. The proprietary format is to secure sensitive data while keeping the invaluable flexibility of microdata access. A web service is also available and benefits national organizations reluctant to give away data but is ready to provide public access to data and/or provide privileged access to selected users. The program is freely available via the Internet. REDATAM has been developed over the last two decades thanks to the financial support from several international organizations (ECLAC-United Nations, UNFPA, the Canadian Government through CIDA and IDRC agencies, IDB and others).
(http://www.cepal.cl/celade)

Countries with 1990 round censuses in REDATAM:

Latin America: Argentina, Brazil, Chile, Colombia, Dominican Republic, Guatemala, Honduras, Nicaragua, Paraguay, Suriname, Uruguay, Venezuela, and English-speaking Caribbean.
Asia: Cambodia* and North Korea*
Africa: Benin, Burkina Faso, Burundi, Cameron, Egypt, Gabon, Ghana, Kenya, Madagascar, Mali, Nigeria, Rwanda*, Seychelles, and Zimbabwe*.

* = database with 100% of the microdata for the population

While these microdata files are not in an archival format in the strict sense, they have been captured in a way that allows for the authoring agency to output a formatted ASCII file with complete structural metadata physically encapsulated to ensure future understandability. It is important that formats such as REDATAM not be viewed as long-term archival formats. The problem with not creating an archival copy and maintaining records in a proprietary format is the cost of eventually having to migrate that information to another format. Proprietary formats soon become legacy formats that due to age, dependency on legacy languages, systems, or hardware becomes difficult, costly and sometimes impossible to migrate.

3.2 Preserving the process

Several manuals and handbooks on performing and managing a national census give detailed lists of procedures and processes. This type of information and details of particular approaches and methodologies are needed for accurately interpreting the resulting data. In addition to this information, consideration of the types of process information that will be of value in preserving institutional memory is useful and often missed. This involves recording and preserving the why as well as the how of the census process. Capturing this information as decisions are made is more cost-effective than reconstructing it at a later date. Attention should be paid to capturing it in a non-proprietary format to reduce the likelihood that the information will be lost due to migration costs.

The complete census cycle consists of four phases (United Nations, Department of Economic and Social Affairs, Statistical Division. *Handbook on Census Management)*:

- Preparation
- Field Operations
- Data Processing
- Evaluation

For each phase, the following documentation is of particular interest:

- reports on procedures and methods.
- comparison of concepts and procedures with the preceding census and current international standards
- evaluation reports for each cycle of the census and the most important documents on which the reports are based
- record books (from manager of the census to the enumerator's logbook) although these may be too raw for general dissemination

The final step is to document the data disseminated and the associated documentation and notes.

4.0 How to assess future value

4.1 Who are the stakeholders

As is clear from the above discussion, standards for census microdata and microdata access are still emerging. The major question of preserving and allowing access to microdata is no longer a technical one but a question of policy. As access to and use of census data expands, the character and complexity of stakeholder groups also expands. These stakeholder groups will be found among governmental, non-governmental, academic, commercial and new user groups. While there will always be competing interests for what should be retained, common interests among several groups will help to identify materials for long-term preservation. Consulting these stakeholder groups early in the process increases the likelihood of obtaining and maintaining funding for long-term preservation.

4.2 Future impact and anticipated use

In terms of the future impact and use patterns for census data, preserving and retaining access to microdata holds the greatest potential. Census data is used to address specific social, economic and demographic issues that change in character over time. The ability to create comparable aggregations over time or for new emerging geographic areas or definitions rest solely on the retention of microdata files. This is particularly true for small area statistics that cannot be derived from larger area aggregations.

For topical tabulations the difficulties are clear. The timing cycle of the census and the production of statistical aggregations mean that the questions asked and the tables created often reflect the concerns and interests from five to ten years prior to publication. New tabulations may no longer be comparable to tabulations from previous censuses due to change in cohort or classification groupings, universe for the table or other changes in definition. Without access to microdata, however it is secured, researchers and analysts are left with few options. In addition, microdata lends itself to research and scholarly discourse and increases the value derived from an individual census.

Consider the example of Canada: In the 1970s National Statistical Services began to disseminate census microdata samples in growing numbers. In Canada, the 1971 revision to the Statistics Act made possible the public release of non-confidential microdata (Tambay and White 2001). Since the 1970s Statistics Canada, with its series of quinquennial enumerations, has regularly issued census microdata samples. Until 1996, researchers had to request samples individually and distribution was highly restricted. In that year a data liberation initiative was instituted to permit Canadian universities to disseminate microdata samples to researchers and their students. The result was an explosion of research. While before liberation five or ten scholars might acquire microdata samples per year, afterward, a single sample at a major university might be accessed hundreds of times per month. The profusion of suppliers means that usage statistics are now impossible to compile, where before the agency recorded every user by name. Given the widespread use of census microdata in the university classroom, Canadian scholars are educating a younger generation of citizens about the utility of the census and democratizing access to census data (Lisa Dillon, private communication, April 21, 2001).

In the United Kingdom, public use census samples called SARs (Samples of Anonymized Records) were first constructed for the 1991 enumeration, with a sample density of 2.0% for individual records, and 0.5% for households. Administrative units with fewer than 120,000 inhabitants were not identified. Notwithstanding the small density of the samples and the absence of geographical detail there was an explosion of research using the SARs. Hundreds of studies were published within six years of the initial release of the data. In anticipation of the 2001 enumeration, disclosure risks were re-assessed, now taking into account error and coding variability as well as differences in timing and coding schemes between datasets. With the permission of the Office of National Statistics privileged access was granted to attempt to match survey records against the SARs (Dale and Elliot, in press). The purpose was to test the practical, as opposed to the theoretical, risks of identifying individuals by matching two sources. The authors reasoned that prior assessments of the likelihood of identifying individuals exaggerated the risks because they neglected to take into account error, differences in timing and incompatibilities of coding schemes. From this rigorous exercise in sleuthing Dale and Elliot conclude:

> For a user of an outside database, attempting this sort of match with no opportunity for verification would prove fruitless. In the first place, the small degree of expected overlap would be a considerable deterrent to an intruder. However, if a match between the two files was attempted the large number of apparent matches would be highly confusing as an intruder would have no way of checking correct identification.

4.3 Informing future censuses

The availability of process information specific to a country and/or organizational structure can inform preparations for future censuses by providing a clear picture of what took place, how it took place, why it was handled in a particular manner, and the successes and difficulties that occurred. This type of information is particularly important for countries with no permanent office or with minimal permanent staff that must essentially create a new system with each census. Providing documentation of why processes and procedures were followed in a specific way is also helpful to international technical consultants, providing them with a clear and well-rounded picture of the previous census activities.

5.0 Inventory of available technology/personnel/knowledge

Prior to the 1990 census round, the United Nations Statistical Division distributed a questionnaire concerning general coverage of the census, organizational structure, cartographic work, house and/or household listing, testing (pre-tests, pilots, etc), the census questionnaires, enumerators and supervisors, enumeration, sampling, data processing, evaluation and analysis, data dissemination, costs and future activities. In addition to the information already requested, the following areas of information related to long-term preservation would be useful in helping countries integrate the discussion of preservation early in the process. Early recognition of preservation needs and possibilities will help in making informed preservation decisions.

- Presence of a national archive capable of preserving digital materials
- Existence of a preservation plan including a record retention schedule
- Types of materials (data products, metadata, process documents) currently preserved and planned for preservation
- Availability of trained preservation staff
- Use of external depositories/archives for data and documentation

6.0 Conclusion

If census microdata are to become widely used, issues of statistical confidentiality must be resolved to the satisfaction of the national statistical agencies and the public as well as researchers. Eurostat sponsored five international conferences on the subject over the past decade. Thanks in part to these efforts and others, the standard practice is now to prepare microdata samples for a variety of users. Among the 52 member-states in the International Monetary Fund's General Data Dissemination System, almost three of every four disseminate census microdata samples, in one guise or another. The development of international microdata standards will increase further the availability of census samples, thereby facilitating comparative research, both in time and space. Everywhere that public dissemination policies have been adopted, an explosion in research has resulted, without a single instance of a breach, or even the allegation of a breach, in statistical confidentiality.

Understanding and incorporating this concept of preservation is important in that it ensures that census data will be protected from loss, alteration, and deterioration. "In this regard , the obligation of archivists is to explain to computer specialists, information technology specialists, and others who are unfamiliar with archives the importance of a physical or logical space, 'independent of the production environment,' where records are protected from loss, alteration, and deterioration so that they may be used as trustworthy evidence as far into the future as is necessary. This is what archiving should be about." (Dollar, 26)

References

Dale, Angela and Mark Elliot. In press. "Proposals for 2001 SARS: An assessment of disclosure risk," *Journal of the Royal Statistical Society, Series A.*

Dollar, Charles M. *Authentic Electronic Records: Strategies for Long-Term Access*. (Chicago, IL:Cohasset Associates, Inc., 2000)

Mexico. Instituto Nacional de Estadística, Geografía e Informática. *Contar 2000. Sistema para la consulta de tabulados y base de datos de la muestra: XII Censo General de Población y Vivienda 2000*. (Aguascalientes, Mexico: 2001).

Ruggles, Steven, Catherine A. Fitch, Patricia Kelly Hall, Matthew Sobek. 2000. "IPUMS-USA: Integrated Public Use Microdata Series for the United States," in Patricia

Kelly-Hall, Robert McCaa and Gunnar Thorvaldsen, eds., *Handbook of International Historical Microdata for Population Research*, Minneapolis MN, 259-284.

Ruggles, Steven, J. David Hacker, and Matthew Sobek. 1995. “Order out of chaos: General design of the Integrated Public Use Microdata Series.” *Historical Methods* 28: 33-39.

Secretariat. 2001. "Report of the March 2001 Work Session on Statistical Data Confidentiality," Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Skopje.

Tambay, Jean-Louis and Pamela White. 2001. "Providing greater accessibility to survey data for analysis," Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Skopje March.

United Nations, Department of Economic and Social Affairs, Statistics Division. *Handbook on Census Management for Population and Housing Censuses, Studies in Methods, Series F No. 83*, (New York: United Nations, 2000)

United Nations, Department of Economic and Social Affairs, Statistics Division. *Handbook Population and Housing Censuses: Part 1 Planning, Organization and Administration of Population and Housing Censuses, Studies in Methods, Series F No. 54*, (New York: United Nations, 1992)

United Nations, Department of Economic and Social Affairs, Statistics Division. *Handbook Population and Housing Censuses: Part 2 Demographic and Social Characteristics, Studies in Methods, Series F No. 54*, (New York: United Nations, 1992)

United Nations, Department of International Economic and Social Affairs, Statistics Division. *Manual on Population Census Data Processing using Microcomputers, Studies in Methods, Series F No. 53*, (New York: United Nations, 1990)

United Nations, Department of International Economic and Social Affairs, Statistics Division. *Emerging Trends and Issues in Population and Housing Censuses, Studies in Methods, Series F No. 52*, (New York: United Nations, 1991)

Vietnam. General Statistics Office. *Data and results from the 3% sample of The Population and Housing Census*. (Hanoi: Central Data Processing Centre, August 2000).