

PRESERVING CENSUS MICRODATA AND MAKING THEM USEFUL: SUDAN

Paper prepared for the First Arab Statistical Conference
Amman, Jordan
November 12-13, 2007

Prof. Awad Hag Ali (Central Bureau of Statistics, Sudan)
and Prof. Robert McCaa (University of Minnesota Population Center)

Research for this paper was funded in part by the National Science Foundation (USA),
grant SES-0433654 'International Integrated Microdata Series'.

Abstract. Census microdata (computerized individualized records from census questionnaires) constitute the most valuable statistical treasure of a nation. Nevertheless, some national statistical offices still do not archive these riches and even fewer make microdata available for research. This paper describes a joint census microdata preservation and dissemination project between the Central Bureau of Statistics of Sudan and the University of Minnesota Population Center IPUMS project. The IPUMS-Sudan project, part of a global initiative funded by the National Science Foundation (USA), has five goals. First we recover the microdata of the 1973 census with the assistance of a commercial data recovery firm. Second, we construct nationally representative, high precision, anonymized samples for the censuses of 1973, 1983 and 1993. Third, we integrate these samples into an international database (<http://international.ipums.org>), along with microdata for more than 30 other countries, including, among the Arab States, Egypt, Iraq, and Palestine. Fourth, we make extracts of samples available to approved researchers without charge. Fifth, we train official statisticians, academics, and policy researchers in the use of the database for comparative research over time and between nations. The IPUMS project is cited as a case study of “good practices” for managing access to census microdata by the UNECE Joint Committee on Microdata Access. Statistical agencies of the Arab States are cordially invited to consider participation in the IPUMS global initiative.

Figure 1. The Problem: Census Microdata at Risk
Computer tapes of Sudan in situ



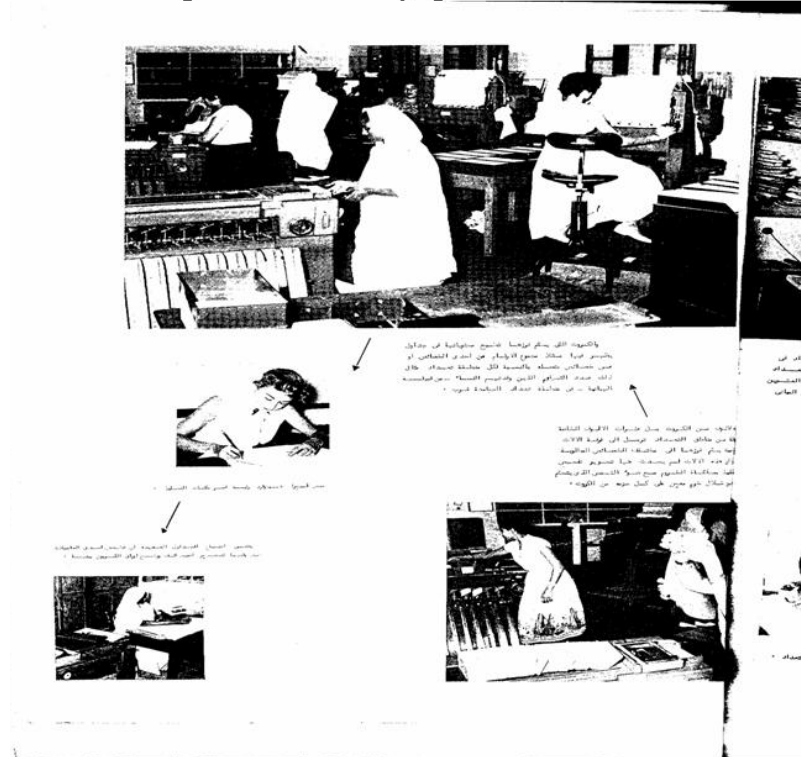
Introduction. Four years ago (September 8-10, 2003), at the “Statistical Capacity Building for the Arab Region” conference in Amman, Jordan, an invitation to participate in the IPUMS-International census microdata project was extended to the region’s official statistical agencies by the American co-author of this paper. Here, we report on a project for a single country, Sudan, to recover and preserve the census microdata held by the Central Bureau of Statistics.

Census microdata. Census *microdata* provide information about individual persons and often families, households, and dwellings, usually in the form of one or more records per case, each consisting of a series of variables. Typical census microdata variables for person records include age, sex, marital status, family relationship, country of birth, educational attainment, employment status, etc. Microdata are exceedingly useful because they allow researchers to interrelate any desired set of population and housing characteristics (Dale, Fieldhouse and Holdsworth, 2000). The flexibility offered by microdata is essential for comparative cross-national as well as chronological research because aggregate tabulations produced by national statistical offices are usually not comparable from one country or census to another. Moreover, even where microdata for several censuses are made available by a national statistical office, rarely are efforts made to standardize sampling, anonymization or coding between censuses. Nor is much guidance offered to researchers seeking to analyze more than one sample. Nevertheless, in the few countries where census microdata covering multiple census years have been easily available to researchers, these data are the most widely-used source for the study of large-scale economic and demographic transformations (McCaa and Ruggles, 2002).

Remarkably, the United Nations Statistics Division (UNSD) has long remained silent with respect to the archiving or preservation of census microdata. Thus, for the 2000 round of population and housing censuses its Principles and Recommendations (Revision 1, 1998) provides little advice. Revision 2 (2007) offers guidance regarding both preserving and disseminating microdata for the upcoming round of censuses, however there is no recommendation regarding archiving or preserving microdata for historical censuses. Given the enormous cost of conducting a census, it would seem wasteful to not expend a small fraction for the conservation of such valuable national resources.

Censuses of Sudan. The first national census of Sudan, conducted in 1956 by British colonial authorities, reported a total population of 10,262,536. The 1956 census is a landmark not only for all subsequent censuses of Sudan, but indeed in the global history of census-taking (see Figure 2). The 1956 census was the first African census to use up-to-date data processing techniques (key-punched cards). The entire census operation was wholly modern in design, execution, and evaluation. The entire nation was enumerated, including nomads, thanks to the extraordinary cooperation of the public, tribal chieftains, and local as well as regional and national authorities. No one was excluded, irrespective of religion, tribe or region. The keyed cards, which in fact constitute census microdata, were preserved for over 25 years. Unfortunately when the central statistical offices in Khartoum were re-located in the 1980s, a decision was made to re-cycle the cards and, in doing so, the microdata of this extraordinary census were destroyed.

Figure 2. The 1956 Census of Sudan: A landmark in the History of Census-Taking
 1956: First National Census of Sudan:
 processed by punch cards



Beginning in 1973, three successive decennial censuses of Sudan were conducted, reporting populations of 14,113,590, 20,594,197 and 24,941,000 for 1973, 1983 and 1993, respectively. These censuses used uniform methodologies. While efforts were made to enumerate the entire population by name, for the rural population only basic characteristics using a short form were collected. A long form was administered to the entire urban population, but only a sample was made of rural districts.

Recovering the census microdata of Sudan. As Figure 1 indicates, when this picture was taken February 11, 2005, the microdata of the censuses of Sudan were at grave risk. At the invitation of the Director General of the Central Bureau of Statistics, the Projects Coordinator of the IPUMS-International initiative flew to Khartoum to assess the likelihood of recovery. Three tapes, randomly selected as a test, were hand-carried to the United States and entrusted to Muller Media Inc., a data recovery company experienced in highly confidential undertakings for governments worldwide, including the Government of Qatar.

With the successful recovery of bits and bytes on those tapes, the CBS approved the recovery of 32 tapes for the 1973 census. Due to the economic blockage by the United States Treasury and the Department of Homeland Security, the project was delayed for more than a year. Finally, the National Science Foundation approved a “work-around” involving a university in Spain which acts as a “go-between” for the project.

Figure 3. Shipment of tapes as received by the data recovery company in New York
1973 Census tapes of Sudan recovered by IPUMS



In August 2007, the tapes were received by the data recovery firm (see Figure 3). Within a matter of days, all the data on one-fourth of the tapes were fully recovered without any special treatment. The recovery effort continues on the remainder. Some tapes must be re-spooled. Others must be manually spaced beyond bad areas to overcome crumpled or unreadable segments of tape (Figure 4). At this point it is unclear what the total fraction of data will be recovered. To date data recovery costs to the project total US\$1,500.

Figure 4. Census tape requiring manual spacing if the data are to be recovered
1973 census tape #1: manual spacing required to recover data



Muller Media has had great success in recovering census data from the 1970s for the IPUMS project. In the case of the 1977 census of Romania, the recovery rate was 97% compared with 95% for the 1976 census of Mali, but only 68% for the 1979 census of Kenya. In each case, copies of the data were repatriated to the National Statistical Office-owner, which was then invited to examine the consistency and comprehensiveness of the recovered data. Happily, in each case, the NSO approved entrusting copies of the microdata for inclusion in the harmonization and dissemination phases of the IPUMS-International project.

Reconstructing census datasets. Once data are recovered, a dataset must be reconstructed, including the preparation of a data dictionary as well as the certification of geographic areas recovered. For the 1973 census, a data dictionary is in the possession of the University of Pennsylvania. Although the dictionary does not match the original raw data being recovered, it is a good starting point. Hopefully, in the not too distant future, the custodian of the dictionary, Dr. Tukufu Zuberi, will repatriate a copy to the Central Bureau of Statistics to assist in the recovery effort.

The 1983 and 1993 census microdata are fully recovered, but the files require additional processing to construct validated datasets. Once this step is completed samples will be drawn of the long forms using, where possible, the standard IPUMS design of every tenth dwelling, following a random start. Expansion factors will be computed and attached to the records so that the age and sex structure of the sampled population will match the published totals for the lowest administrative units identified in the samples. Dictionaries and codebooks must also be verified against the data to be sure that all codes and labels are correct.

Table 1 near here

For the Arab region as a whole, the status of microdata for historical censuses is not well known (see Table 1). Delegates to the conference are invited to examine Table 1 and email corrections or suggestions to Prof. Robert McCaa (rmccaa@umn.edu).

Beyond data recovery: integrated, anonymized scientific samples for academic and policy research. The IPUMS-International project, not only preserves census microdata and documentation, it also democratizes access to anonymized, integrated census samples. This global initiative transforms raw census microdata into confidentialized, integrated high-density samples accessible to researchers and policy makers everywhere (Ruggles et. al. 2003). In almost all cases, new samples are being drawn according to uniform specifications developed by the IPUMS project. For private households, every n^{th} (where $n = 10$ for a 10% sample) household is selected after a random start from a dataset where the households are ordered geographically down to the census tract level. For collective or group quarters, the sample unit is the individual; that is, after a random start, every n^{th} person is selected (McCaa and Esteve 2006).

Begun in 1999 with funding provided by the National Institutes of Health and the National Science Foundation of the United States, to date the IPUMS initiative enjoys official endorsement by statistical agencies in more than seventy countries, encompassing almost two-thirds of the world's population (Table 2). Similar to the National Science Foundation funded Hubble telescope project, to which IPUMS is sometimes compared, the IPUMS-International web-site facilitates access to researchers world-wide, regardless

of country of birth, residence, or citizenship. Although access is restricted to bona fide researchers, no cost is charged for access.

Table 2 near here

In May 2002, the first phase of IPUMS international integrated census microdata were made available to researchers for Colombia (1964-1993), France (1962-1990), Kenya (1989-1999), Mexico (1960-2000), the United States (1960-2000), and Vietnam (1989-1999), followed in 2003 by China (1982) and in 2004 by Brazil (1960, 1970, 1980, 1991, 2000). In 2007, samples from 26 countries are available. Currently the IPUMS-International website offers more than 200 million person records consisting of hundreds of variables from samples for 80 censuses (Table 3). Over the next five years, thanks to sustained funding by the National Science Foundation and the National Institutes of Health, microdata for 20-30 more countries will be added through regional initiatives in Europe, Latin America, Asia, Africa and the Arab region.

Table 3 near here

Some of the most respected statistical organizations in the world are cooperating in the project, such as INSEE-France, Statistics Canada, Office of National Statistics (UK), Federal Statistics Office (Germany), National Bureau of Statistics (China), the Census Bureau of the United States, etc. Most National Statistical Office partners take advantage of the opportunity to draw new high precision, household samples according to IPUMS specifications, with costs shared by the project. Among the Arab States, a 10% sample of the 1997 census of Palestine is currently available from the project website. In 2008, anonymized, integrated samples of Egypt and Iraq will be launched, thanks to the official endorsement of the IPUMS project by the Central Agency for Public Mobilisation and Statistics (CAPMAS) and the Central Bureau of Statistics (CBS).¹ In 2009, the recovery and integration of samples for Sudan will be launched.

“Good Practice” for managing access to census microdata. It is important to understand that the IPUMS project does not simply receive raw census microdata on CDs to pass copies around to researchers. On the contrary, once official permission is granted to the project by the corresponding national statistical authority, typically three years of work are invested in transforming the raw data into confidentialized, integrated samples. These labors have been successful, as recognized by the United Nations Economic Commission for Europe. IPUMS is the only academic initiative cited by the UNECE joint task force on Managing Statistical Confidentiality and Microdata Access: Principles and Guidelines of Good Practice (2007 <http://www.unece.org/stats/documents/tfcm.htm>; Annex 1.23).

Invitation. National Statistical Offices of the Arab Region are cordially invited to become associated with the project, to facilitate the recovery of data for historical censuses and to assist in expanding the database by endorsing this global initiative. Now that the construction of anonymized microdata data samples is becoming a reality, integration of census microdata is an obvious next step to enhance use. With the emergence of global standards for harmonizing census data spurred by the Statistics Division of the United Nations and the increase in power of ordinary desktop computers, the major challenge that remains is the actual construction of integrated census microdata

¹ Regrettably, following the invasion of Iraq orchestrated by President George W. Bush in March 2003, the entire database for the 1987 census was destroyed in the looting of CBS facilities.

samples. Thanks to the cooperation of some 70 official census agencies worldwide and with the financial support of the National Science Foundation and the National Institutes of Health, the IPUMS-International project is committed to anonymizing and integrating microdata for 150 censuses by 2010. If the IPUMS-International project is truly successful it will continue beyond the 2000 round of censuses, incorporating samples of participating countries for the 2010 censuses as they become available. The number of users and stakeholders will increase proportionately.

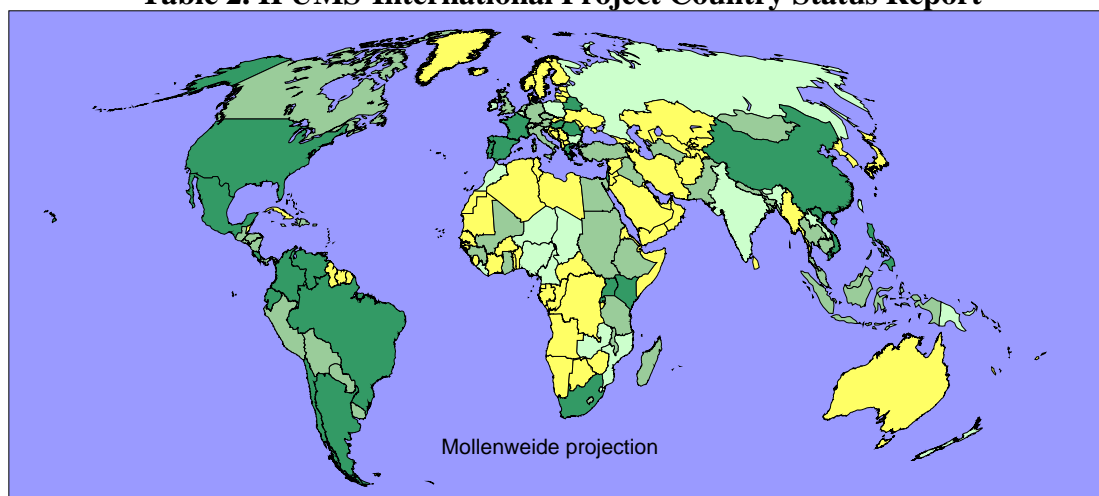
References.

- Dale, A., E. Fieldhouse, and C. Holdsworth. 2000. *Analyzing census microdata*. Arnold: London.
- McCaa, Robert and Albert Esteve. 2006. "IPUMS-Europe: Confidentiality measures for licensing and disseminating restricted access census microdata extracts to academic users," *Monographs of official statistics: Work session on statistical data confidentiality*. Luxembourg: Office for Official Publications of the European Communities, pp. 37-46.
- McCaa, Robert and Steven Ruggles. 2002. The Census in Global Perspective and the Coming Microdata Revolution. In Vol. 13, *Nordic Demography: Trends and Differentials, Scandinavian Population Studies*, edited by J. Carling. Oslo: Unipub/Nordic Demographic Society, pp. 7-30.
- McCaa, Robert and Wendy Thomas. 2003. "Archiving Census Documentation and Microdata: Preserving Memory, Increasing Stakeholders, *Notas de Población* XXIX:75:303-320; coauthor: Wendy L. Thomas. "[Symposium on Global Review of 2000 Round of Population and Housing Censuses: Mid-Decade Assessment and Future Prospects. Statistics Division, United Nations Secretariat, New York, Aug. 7-10, 2001].
- United Nations Statistics Division. (2007). *Principles and Recommendations for Population and Housing Censuses*. Revision 2. Department of Economic and Social Affairs, New York.

Table 1. Arab States: Census Microdata Inventory by Year and Country
Bold Country = year of (expected) launch of anonymized samples by IPUMS project
Bold = microdata exist; ** = data entrusted to project;
italics = data were lost or not computerized; "?" existence of microdata not known
Please correct the entry in this table for your country

| Country | 1970s | 1980s | 1990s | 2000s |
|---------------------------------------|---------------|---------------|---------------|-------------|
| Algeria | 1977? | 1987? | 1998 | . |
| Bahrain | 1971? | 1981? | 1991 | 2001 |
| Djibouti | | 1983? | | |
| Egypt (to be launched in 2008) | <i>1976</i> | 1986** | 1996** | 2006 |
| Iraq (to be launched in 2008) | <i>1977</i> | <i>1987</i> | 1997** | |
| Jordan | 1972? | 1983? | 1995 | 2004 |
| Kuwait | 1970? | 1980, 85? | 1995 | ... |
| Lebanon | 1970? | | | |
| Libya Arab Jamahiriya | 1973? | 1984? | | 2000 |
| Mauritania | 1977? | 1988? | | 2000 |
| Morocco | 1971? | 1982? | 1994 | 2002 |
| Oman | | | 1993 | 2003 |
| Palestine (launched in 2007) | | | 1997 | . |
| Qatar | | 1986? | 1997 | 2004 |
| Saudi Arabia | 1974? | | 1992 | 2004 |
| Sudan (to be launched in 2009) | 1973** | 1983** | 1993** | 2007 |
| Syria | 1970? | 1981? | 1994 | 2004 |
| Tunisia | 1975? | 1984? | 1994 | 2004 |
| United Arab Emirates | 1975? | 1980, 85? | 1995 | 2005 |
| Yemen | 1973? | 1983? | 1994 | 2004 |
| Total Censuses | 16 | 16 | 16 | 14 |
| Census Microdata Archived | 1 | 2 | 16 | 14 |

Table 2. IPUMS-International Project Country Status Report



Key: dark green = disseminating; medium green = processing data; light green = negotiating MOU

| Region | Country |
|-----------------------------------|---|
| Africa 15 countries | Egypt, Ethiopia, Ghana, Guinea (Conakry), Kenya, Lesotho, Madagascar, Malawi, Mali, Mauritius, Rwanda, Sierra Leone, South Africa, Sudan, Uganda |
| Asia 17 countries | Armenia, Bangladesh, Cambodia, China, Fiji Islands, Indonesia, Iraq, Israel, Malaysia, Mongolia, Nepal, Pakistan, Palestine, Philippines, Thailand, Turkmenistan, Vietnam |
| Europe 16 countries | Austria, Belarus, Bulgaria, Czech Republic, France, Germany, Greece, Hungary, Italy, Netherlands, Portugal, Romania, Slovenia, Spain, Turkey, United Kingdom |
| North America 12 countries | Canada, Costa Rica, Dominican Republic, El Salvador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Puerto Rico, Saint Lucia, United States of America |
| South America 10 countries | Argentina, Brazil, Bolivia, Chile, Colombia, Ecuador, Paraguay, Peru, Uruguay, Venezuela |

Table 3. Sample Characteristics (fraction, no of household and person records, etc.) by Country and Census Year

Note: “Bolded Country, Census” indicates preserved with IPUMS assistance

| Census | Sample Fraction (%) | Households | Persons | Weighted |
|-----------------------|---------------------|------------|------------|----------|
| Argentina 1970 | 2 | 129,728 | 466,892 | – |
| Argentina 1980 | 10 | 672,062 | 2,667,714 | yes |
| Argentina 1991 | 10 | 1,148,351 | 4,143,727 | yes |
| Argentina 2001 | 10 | 1,040,852 | 3,626,103 | – |
| Belarus 1999 | 10 | 385,508 | 990,706 | – |
| Brazil 1960 | 5 | 613,273 | 3,001,439 | – |
| Brazil 1970 | 5 | 1,022,207 | 4,953,759 | yes |
| Brazil 1980 | 5 | 1,343,377 | 5,870,467 | yes |
| Brazil 1991 | 5.8 | 2,012,276 | 8,522,740 | yes |
| Brazil 2000 | 6 | 2,652,356 | 10,136,022 | yes |
| Cambodia 1998 | 10 | 223,513 | 1,141,254 | – |
| Chile 1960 | 1 | n.a.* | 88,184 | – |
| Chile 1970 | 10 | 199,041 | 890,481 | – |
| Chile 1982 | 10 | 282,356 | 1,133,062 | – |
| Chile 1992 | 10 | 373,964 | 1,335,055 | – |
| Chile 2002 | 10 | 486,115 | 1,513,914 | – |
| China 1982 | 0.1 | 242,718 | 1,002,691 | – |
| Colombia 1964 | 2 | n.a.* | 349,652 | – |
| Colombia 1973 | 10 | 349,853 | 1,988,831 | – |
| Colombia 1985 | 10 | 571,046 | 2,643,125 | yes |
| Colombia 1993 | 10 | 774,321 | 3,213,657 | – |
| Costa Rica 1963 | 5 | n.a.* | 82,345 | – |
| Costa Rica 1973 | 10 | 36,323 | 186,762 | – |
| Costa Rica 1984 | 10 | 56,186 | 241,220 | – |
| Costa Rica 2000 | 10 | 106,973 | 381,500 | – |
| Ecuador 1962 | 3 | n.a.* | 136,443 | – |
| Ecuador 1974 | 10 | 145,902 | 648,678 | yes |
| Ecuador 1982 | 10 | 195,401 | 806,834 | – |
| Ecuador 1990 | 10 | 243,898 | 966,234 | – |
| Ecuador 2001 | 10 | 354,222 | 1,213,725 | – |
| France 1962 | 5 | 748,917 | 2,320,901 | – |
| France 1968 | 5 | 815,699 | 2,487,778 | – |
| France 1975 | 5 | 915,624 | 2,629,456 | – |
| France 1982 | 5 | 969,632 | 2,631,713 | – |
| France 1990 | 4.2 | 949,893 | 2,360,854 | – |
| Greece 1971 | 10 | 249,350 | 845,483 | – |
| Greece 1981 | 10 | 294,323 | 923,108 | – |

| | | | | |
|---|------|------------|-------------|-----|
| Greece 1991 | 10 | 320,387 | 951,875 | – |
| Greece 2001 | 10 | 367,438 | 1,028,884 | – |
| Hungary 1970 | 5 | 172,831 | 515,119 | – |
| Hungary 1980 | 5 | 211,355 | 536,007 | – |
| Hungary 1990 | 5 | 219,389 | 518,240 | – |
| Hungary 2001 | 5 | 227,252 | 510,502 | – |
| Israel 1972 | 10 | 89,190 | 315,608 | – |
| Israel 1983 | 10 | 124,610 | 403,474 | – |
| Israel 1995 | 10 | 177,412 | 556,365 | – |
| Kenya 1989 | 5 | 224,861 | 1,074,098 | – |
| Kenya 1999 | 5 | 317,106 | 1,407,547 | – |
| Mexico 1960 | 1.5 | n.a.* | 502,800 | – |
| Mexico 1970 | 1 | 82,856 | 483,405 | – |
| Mexico 1990 | 10 | 1,648,280 | 8,118,242 | – |
| Mexico 2000 | 10.6 | 2,312,035 | 10,099,182 | yes |
| Palestine 1997 | 10 | 40,753 | 259,191 | yes |
| Philippines 1990 | 10 | 1,156,126 | 6,013,913 | yes |
| Philippines 1995 | 10 | 1,362,190 | 6,864,758 | – |
| Philippines 2000 | 10 | 1,511,890 | 7,417,810 | yes |
| Portugal 1981 | 5 | 179,409 | 492,289 | – |
| Portugal 1991 | 5 | 214,155 | 491,755 | – |
| Portugal 2001 | 5 | 258,843 | 517,026 | – |
| Romania 1992 | 10 | 728,846 | 2,238,578 | – |
| Romania 2002 | 10 | 732,016 | 2,137,967 | – |
| Rwanda 1991 | 10 | 153,041 | 742,918 | – |
| Rwanda 2002 | 10 | 191,719 | 843,392 | – |
| South Africa 1996 | 10 | 993,801 | 3,621,164 | yes |
| South Africa 2001 | 10 | 991,543 | 3,725,655 | yes |
| Spain 1981 | 5 | n.a.* | 2,084,221 | yes |
| Spain 1991 | 5 | 592,276 | 1,931,458 | yes |
| Spain 2001 | 5 | 714,473 | 2,039,274 | – |
| Uganda 1991 | 10 | 339,166 | 1,548,460 | yes |
| Uganda 2002 | 10 | 529,271 | 2,497,449 | – |
| United States 1960 | 1 | 579,212 | 1,799,888 | – |
| United States 1970 | 1 | 744,475 | 2,029,666 | – |
| United States 1980 | 5 | 4,711,341 | 11,343,120 | – |
| United States 1990 | 5 | 5,527,406 | 12,501,046 | yes |
| United States 2000 | 5 | 6,184,438 | 14,081,466 | yes |
| Venezuela 1971 | 10 | 284,336 | 1,158,527 | yes |
| Venezuela 1981 | 10 | 323,321 | 1,441,266 | – |
| Venezuela 1990 | 10 | 468,808 | 1,803,953 | yes |
| Vietnam 1989 | 5 | 534,223 | 2,626,985 | yes |
| Vietnam 1999 | 3 | 534,139 | 2,368,167 | yes |
| TOTAL | | 57,681,479 | 202,185,219 | |
| * Sample is not organized into households. | | | | |
| Source: http://international.ipums.org/international/sample_summary.html | | | | |

IPUMS Case Study appended to UNECE **Managing Statistical Confidentiality and Microdata Access: Principles and Guidelines of Good Practice (2007)**: <http://www.unece.org/stats/documents/tfcm.htm>

Annex 1.23: Case Study – Arrangements for Providing International and National Access to Anonymized Census Microdata Samples via the IPUMS-International and the Integrated European Census Microdata websites (University of Minnesota Population Center and the Centre d'Estudis Demogràfics, Autonomous University of Barcelona) with France, as a specific example. January 1, 2007.

1. Broad description

High precision, anonymized, integrated census microdata are available to researchers on a restricted access basis from IPUMS-International (www.ipums.org/international). Terms are specified by a memorandum of understanding negotiated between each National Statistical Office and the University of Minnesota. This method of dissemination is governed, on the one hand, by legislation requiring that the data be held in strict confidence and used exclusively for statistical purposes and, on the other, by a stringent, internationally enforceable license agreement between the University of Minnesota and each user. In May 2002, anonymized, integrated microdata samples for the French censuses of 1962, 1968, 1975, 1982 and 1990 were released, along with samples for China, Colombia, Kenya, Mexico, the USA and Vietnam. The December 2006 release includes samples for the censuses of Belarus, Greece, Romania and Spain as well as the Philippines and Uganda. As of January 1, 2007, the database comprises 63 samples, 20 countries, and 185 million person records. An additional six European statistical agencies (and 38 non-European) have provided census microdata to the project: Austria (4 censuses), Czech Republic (2), Hungary (4), Netherlands (3), Portugal (3), and the United Kingdom (2); the 1981 and earlier censuses are under consideration). Four other European countries have endorsed the project, but have not yet provided data: Bulgaria, Germany, Italy, and Slovenia. The European microdata will also be distributed by the Integrated European Census Microdata (IECM) project using identical protocols, although the microdata will be harmonized according to European, rather than global, practices.

2. Why is it a good practice?

Conditions of access are transparent and provide a degree of certainty to users and the National Statistical Offices. Sanctions for violations of misuse are clearly spelled out and enforceable by a set of strong administrative and legal mechanisms. The microdata are anonymized by means of a variety of technical measures, including the suppression of detailed geography. Variables are integrated using a composite coding scheme to facilitate temporal and cross-national comparative research. The documentation, including both scanned images of forms and instructions as well as integrated metadata, is extensive and available at no cost. The microdata are also available at no cost, but availability is restricted to approved academic and policy researchers. These practices are in compliance with the Fundamental Principles of Official Statistics.

3. Target audience

The research community, including academic and policy makers regardless of country of birth, residence, work-place or citizenship.

4. Detailed description

The IPUMS-International project is governed by a uniform Memorandum of Understanding (MOU) signed with each participating National Statistical Office. The MOU (copy appended below) confirms that the National Statistical Office specifies the terms and conditions under which the microdata and metadata entrusted to the University of Minnesota and the Autonomous University of Barcelona shall be governed:

- 1) the NSO retains ownership, including copyright;
- 2) data are to be used for purposes of teaching, research, and publishing;
- 3) use for commercial or income generating purposes is prohibited;
- 4) application procedures for obtaining access to microdata are specified in the MOU;
- 5) confidentiality of the data is protected by means of prohibitions against
 - a. any attempt to ascertain the identity of individuals, families, households, dwellings or other identities
 - b. any allegation that an identification has been made.

In addition there are statements regarding:

- 6) the necessity of security measures for retaining microdata;
- 7) publication and citation requirements;
- 8) procedure for dealing with violations, including sanctions;
- 9) the sharing of integrated microdata with the National Statistical Offices;
- 10) recognition of jurisdiction under international law with the ICC International Court of Arbitration for the settlement of disputes; and
- 11) establishing the supreme precedence of the MOU over any subsidiary document, contract or other instrument.

The sanctions clause of the MOU, which is particularly important for assuring compliance, reads:

“Violation of the user license may lead to professional censure, loss of employment, and/or civil prosecution. The University of Minnesota, national and international scientific organizations, and the [the Statistical Agency of Country X] will assist in the enforcement of provisions of this accord.”

4.1 Data confidentiality

Before providing census microdata to the Minnesota Population Center, the National Statistical Office imposes a

number of undisclosed technical confidentiality measures. The Minnesota Population Center imposes an additional suite of techniques such that any allegation that an individual has been identified with absolute certainty is false. In addition, to further ensure the confidentiality of the microdata, administrative geography is limited. In the case of France 22 regions are identified. The smallest has a population exceeding 80,000 in the 1990 census (sample n > 4,000). The sample count for any identifiable single year of age is >100. For any identifiable country of citizenship the sample count is >100. Each National Statistical Office determines the minimum population threshold for the identification of administrative geography and other sensitive characteristics, such as ethnicity, country of birth, citizenship, etc.

4.2 Rules and procedures regarding release to users

Prospective users must complete an electronic application to gain access to the data. The preamble of the application reads:

Legal notice: Submission of this application constitutes a legally binding agreement between the applicant, the applicant's institution, the University of Minnesota, and the relevant official statistical authorities. Submitting false, misleading or fraudulent information constitutes a violation of this agreement. Misusing the data by violating any of the conditions detailed below also constitutes a violation of this agreement and may lead to professional censure, loss of employment, or civil prosecution under relevant national and international laws, and to sanctions against your institution, at the discretion of the University of Minnesota and the official statistical authorities."

The application form requires that the applicant indicate agreement, by electronically checking specifically each of eight conditions of use, including the following:

 **Use of the microdata must follow strict rules of confidentiality.**

Users will maintain the confidentiality of persons and households. Any attempt to ascertain the identity of persons or households from the microdata is prohibited. Alleging that a person or household has been identified in these data is also prohibited. Statistical results that might reveal the identity of persons or entities may not be reported or published in any form.

And:

 **Any violation of this license agreement will result in disciplinary action, including possible loss of employment.**

Violation of this agreement will lead to revocation of this license, recall of all microdata acquired, a motion of censure to the relevant professional organization(s) and civil prosecution under national or international statutes, at the discretion of the Regents of the University of Minnesota and the official statistical agencies. Sanctions likewise may be taken against the institution with which the violator is affiliated.

Failure to indicate agreement with any one of the conditions automatically disqualifies the applicant for access to the microdata. In addition the successful applicant must provide detailed information on academic qualifications, affiliation, research experience, source of funding, bona fides, and familiarity with human subjects protections regarding statistical confidentiality. Finally the applicant must submit a project description demonstrating need for access to census microdata. Applications are reviewed by senior principal investigators. Approximately 1/3 of applicants who complete the form are denied access. The application is valid for one year and may be renewed.

5. Supporting legislation (example of France)

Article 6 of the law of 1978 introduced the possibility for statisticians and researchers to use personal data, including nominative data, originally collected for purposes other than historical or scientific research or statistics. More precisely, it indicates that subsequent processing for statistical or research purposes is always compatible with the objectives for which the data had been collected. French Act no. 2004-801 of August 6, 2004 amends and updates the Statistics Law of 1978 to protect individuals with regard to the processing of personal data and the free movement of these data. The Act is in compliance with the European directive no. 95/46/CE of October 24, 1995 of the European Parliament and Council. Information on legislation regarding good practices is available at:

<http://unstats.un.org/unsd/goodprac/default.asp> For information on statistical confidentiality, microdata access and privacy, see "Principle 6".

6. Strengths

- a) Fosters maximum uniformity of approach and facilitates greater access to microdata by the research community.
- b) Improves on arrangements for providing access to microdata to the greater satisfaction of both the National Statistical Offices and the research community.
- c) National Statistical Offices cede census microdata to the University of Minnesota for dissemination on a licensed basis to approved researchers. All licensed microdata disseminated by the University of Minnesota Population Center are governed by a uniform Memorandum of Understanding (MOU) between the National Statistical Office and the University. If requested to do so, the University will cease dissemination and return all copies of census microdata in its possession to the corresponding National Statistical Office.
- d) Employees of the University who work with original source data are certified in human subject protections, including the protection of statistical confidentiality. Violations are punishable by termination of employment, and, at the discretion of the University, civil prosecution with a maximum fine of US\$250,000

- and/or three years imprisonment.
- e) The means of gaining access to the microdata are transparent and equitable. They are based on the principle of freedom of scientific inquiry, regardless of country of birth, residence, workplace or citizenship. Decisions to grant access are determined by project principal investigators. Each individual who wishes to work with the microdata is required to be licensed. The license is valid for one year and is renewable. A condition for renewal is the sharing of research findings, which, in turn, are made available to the national statistical offices.
 - f) Microdata are available as extracts on a licensed basis only to researchers who agree to abide by the conditions of use and demonstrate a bona fide research need to access the data. The license constitutes a legally binding undertaking. An attempt to match individuals constitutes a violation of the license agreement and would lead to sanctions against the individual and his/her institution.
 - g) Sanctions for breaches of the license agreement are clearly spelled out. These include:
 - i. sanctions against both the individual and the institution with which the individual is associated (e.g., University, international organization);
 - ii. denial of access would immediately be invoked against the individual and his/her institution and would continue until corrective measures were deemed to be sufficient by the University of Minnesota and the National Statistical Office whose data were violated. If the institution where the breach occurred was the recipient of a grant from the National Institutes of Health of the United States, each researcher at the institution could be required to undergo Human Subjects Protection training and re-certification before access was re-instituted for individuals at that institution.
 - iii. civil prosecution could be instituted with assistance requested, under the terms of the project MOU, of the National Statistical Office of the country in which the violation occurred.
 - h) Microdata are encrypted during transmission using 128-bit SSL (Secure Sockets Layer) encryption standards used by the financial industry.
 - i) Anonymization protocols (top coding, bottom coding, grouping of small cell counts, collapsing of variables, randomization of records and some recodes, suppression of sensitive variables, etc.) are rigorous, yet precision of samples is high. Anonymization protocols are determined by each National Statistical Office before extracts of the data are disseminated.
 - j) Integrated metadata are provided describing census operations, sample methodologies, variables and codes. The documentation is harmonized so that researchers who become familiar with the metadata for one census will readily understand the metadata system for any other census of any other country.
 - k) Microdata consist of high precision household samples with many integrated, value-added variables—such as “WTPER”, which specifies the person weight for each record in every sample; “SUBSAMP”, which provides 100 certified sub-samples which researchers may use to generate robust estimates of sample variance; “SPLOC” which points to the spouse of each individual whose spouse in co-resident in a household; etc.
 - l) Costs are borne through sustained funding from the National Science Foundation of the United States of America with supplementary funds provided by the National Institutes of Health. Where required, the project pays a license fee to the National Statistical Office for the documentation and microdata. The fee is intended to cover marginal costs for the National Statistical Office to provide technical assistance in developing the microdata samples and interpreting the documentation. The **European Union Sixth Framework Programme** provides support to the IECM project for enhancing, harmonizing and disseminating the integrated European microdata and metadata as well as for coordinating tasks based in Europe.

7. Weaknesses

- a. National Statistical Offices cede authority to the University to grant access to census microdata extracts to bona fide researchers. Decisions to grant access are determined by project principal investigators.
- b. Microdata are not wholly anonymized. With sufficient resources, in terms of computing power, time, and a companion microdataset, data matching could be performed to identify individuals to a high probability, although not with absolute certainty.
- c. Misuse of microdata by even one researcher may impact negatively on the ability of a National Statistical Office to obtain cooperation of respondents in that country, or even conceivably, other countries.
- d. Microdata extracts are not obtained directly from the National Statistical Office.
- e. Quality of microdata may not be sufficiently high for the intended research purpose.
- f. Whether the license constitutes a legally binding undertaking has not been tested in a court of law.
- g. There is no requirement that the microdata be destroyed once the initial research is completed.
- h. There is no opportunity for the National Statistical Office to comment upon the research before it is published.

8. References

- Bruengger, Heinrich. 2004. “The relationship between the fundamental principle on confidentiality and population censuses: Statement from the UNECE Statistical Division,” *United Nations Symposium on Population and Housing Censuses*: New York, September 13-14.
- Isnard, Michel. 2006. “Statistics and individual liberties: recent changes in French law,” *Courrier des statistiques*, English series no.12, pp. 26-30.

Letter of Understanding (endorsed by more than 70 official statistical agencies
worldwide)

**Integrated Public Use Microdata Series International
and [National Statistics Institute of Country X]**

Purpose. The purpose of this letter is to specify the terms and conditions under which metadata and microdata produced by the [National Statistics Institute of Country X] shall be distributed by **Integrated Public Use Microdata Series International** of the University of Minnesota.

1 **Ownership.** The [National Statistics Institute of Country X] is the owner and licensee of the intellectual property rights (including copyright) in the metadata and microdata of [Country X] acquired by the University of Minnesota to be distributed by **Integrated Public Use Microdata Series International**. This agreement explicitly authorizes release to the University of microdata of [Country X] that may be in the possession of third parties. The University is obligated to provide to the [National Statistics Institute of Country X] timely notice of any such acquisitions and, upon request and without cost, provide copies of same.

2 **Use.** These data are for the exclusive purposes of teaching, scientific research and publishing, and may not be used for any other purposes without the explicit written approval, in advance, of the [National Statistics Institute of Country X].

3 **Authorization.** To access or obtain copies of integrated microdata of [Country X] from **Integrated Public Use Microdata Series International**, a prospective user must first submit an electronic authorization form identifying the user (i.e., principal investigator) by name, electronic address, and institution. The principal investigator must state the purpose of the proposed project and agree to abide by the regulations contained herein. Once a project is approved, a password will be issued and data may be acquired from servers or other electronic dissemination media maintained by **Integrated Public Use Microdata Series International**, the [National Statistics Institute of Country X], or other authorized distributors. Once approved, the user is licensed to acquire integrated metadata and microdata of [Country X] from **Integrated Public Use Microdata Series International** or other authorized distributors. No titles or other rights are conveyed to the user.

4 **Restriction.** Users are prohibited from using data acquired from the **Integrated Public Use Microdata Series International** or other authorized distributors in the pursuit of any commercial or income-generating venture either privately, or otherwise.

5 **Confidentiality.** Users will maintain the absolute confidentiality of persons and households. Any attempt to ascertain the identity of a person, family, household, dwelling, organization, business or other entity from the microdata is strictly prohibited. Alleging that a person or any other entity has been identified in these data is also prohibited.

6 **Security.** Users will implement security measures to prevent unauthorized access to microdata acquired from **Integrated Public Use Microdata Series International** or its partners.

7 **Publication.** The publishing of data and analysis resulting from research using metadata or microdata of [Country X] is permitted in communications such as scholarly papers, journals and the like. The authors of these communications are required to cite [National Statistics Institute of Country X] and **Integrated Public Use Microdata Series International** as the sources of the data of [Country X], and to indicate that the results and views expressed are those of the author/user.

8 **Violations.** Violation of the user license may lead to professional censure, loss of employment, and/or civil prosecution. The University of Minnesota, national and international scientific organizations, and the [National Statistics Institute of Country X] will assist in the enforcement of provisions of this accord.

9 **Sharing.** **Integrated Public Use Microdata Series International** will provide electronic copies to the [National Statistics Institute of Country X] of documentation and data related to its integrated microdata as well as timely reports of authorized users.

10 **Jurisdiction.** Disagreements which may arise shall be settled by means of conciliation, transaction and friendly composition. Should a settlement by these means prove impossible, a Tribunal of Settlement shall be convened which will rule upon the matter under law. This Tribunal shall be composed of an arbitrator, which shall be selected by the ICC International Court of Arbitration. This agreement shall be governed by, and construed in accordance with, generally accepted principles of International Law.

11 **Order of Precedence.** In the event of a conflict between a term or condition of this Letter of Understanding and a term or condition of any Contract, to which this Letter of Understanding is attached, the term or condition in this Letter of Understanding shall prevail.

Date: _____ Signed: _____
Regents of the University of Minnesota By: Kevin J. McKoskey, Sponsored Projects
Administration

Date: _____ Signed: _____
[National Statistics Institute of Country X] By:
Rev. Jan. 27, 2005