Using Census Microdata Disseminated by IPUMS-International to Assess Millennium Development Goals of Literacy, Education and Gender Equity in the Ugandan censuses of 1991 and 2002[1]

Robert McCaa[2], Steven Ruggles[3], and Matt Sobek[4]

Key words: integrated census microdata, IPUMS-International, Millennium Development Goals, literacy, education, gender, Uganda

For good development policy making, democratization of access to integrated census microdata offers important advantages both to prepare development plans as well as assess accomplishments. The IPUMS-International project (www.ipums.org/international) is a global initiative in cooperation with national statistical authorities world-wide to anonymize, integrate and manage access to census microdata. With the IPUMS system, extracts of microdata, adapted to the specific research needs of each user, are distributed as ASCII text-files along with the corresponding metadata without charge via the Internet. Researchers analyze the data using their own software and hardware. To illustrate possible applications with respect to the Millennium Development Goals, microdata from the 1991 and 2002 censuses of Uganda are used to measure progress toward the attainment of universal primary education and the elimination of gender inequities in access to primary education.

"I have recently taken on a new job at [an international organization]
and will be supporting developing countries in carrying out their censuses.
I am therefore interested to understand more about what data
have been collected in the past across countries."
--*application #1759*, www.ipums.org/international

[1] A version of the paper was presented at the Scientific Statistics Conference, Statistics House, Uganda Bureau of Statistics, Kampala, June 12, 2007.

[2] Professor of History, University of Minnesota, Minneapolis, MN 55455 USA. rmccaa@umn.edu
[3] Director, University of Minnesota Population Center, and Professor of History. ruggles@pop.umn.edu
[4] Research Scientist, University of Minnesota Population Center. sobek@pop.umn.edu

Table 1. Extant microdata and datasets entrusted to IPUMS-International project by country and census
bold country = Memorandum of Understanding signed with Regents of the University of Minnesota
Year = census conducted; **Bold year** = microdata survive; m = micro-census;
* = archived by African Census Analysis Project, University of Pennsylvania (Zuberi 2005)

datasets entrusted	Country	2000s	1990s	1980s	1970s	1960s
Accessible from www.ipums.org/international, 1999-2007 80 census samples, 26 countries, 202 million person records						
4	**Argentina**	**2001**	**1991**	**1980**	**1970**	1960
1	**Belarus**		**1999**	1989	1979	1970
5	**Brazil ('60 recovered)**	**2001**	**1991**	**1980**	**1970**	**1960**
1	**Cambodia**		**1998**			
5	**Chile**	**2002**	**1992**	**1982**	**1970**	**1960**
2	**China ('90 coming soon)**	**2000**	**1990**	**1982**		1964
5	**Colombia ('05 coming soon)**	**2005**	**1993**	**1985**	**1973**	**1964**
4	**Costa Rica**	**2000**		**1984**	**1973**	**1963**
5	**Ecuador**	**2001**	**1990**	**1982**	**1974**	**1962**
6	**France ('99 coming soon)**	**1999**	**1990**	**1982**	**1975**	**1968, 2**
4	**Greece ('71 recovered)**	**2001**	**1991**	**1981**	**1971**	1961
4	**Hungary ('70 recovered)**	**2001**	**1990**	**1980**	**1970**	
3	**Israel**		**1995**	**1983**	**1972**	
4	**Kenya ('69 & '79 coming soon)**	**1999**	**1989***	**1979***	**1969***	
5	**Mexico ('80 coming soon)**	**2000**	**1990**	**1980**	**1970**	**1960**
1	**Palestinian Territories**		**1997**			
3	**Portugal**	**2001**	**1991**	**1981**	1970	1960
3	**Romania ('77 coming soon)**	**2001**	**1992**		**1977**	1965
2	**Rwanda**	**2001**	**1991**			
2	**South Africa ('96 and '01 only)**	**2001**	**1996*, 91***	**1985*, 80***	**1970***	**1960**
3	**Spain**	**2001**	**1991**	**1981**	1970	1960
5	**United States**	**2000**	**1990**	**1980**	**1970**	**1960**
2	**Uganda**	**2002**	**1991***	1980*		1969
5	**Venezuela ('01 coming soon)**	**2001**	**1990**	**1981**	**1971**	**1961**
2	**Vietnam ('89 recovered)**		**1999**	**1989**	1979	
Europe						
4	**Austria**	**2001**	**1991**	**1981**	**1971**	1961
	Bulgaria	**2001**	**1992**	**1985**	1975	1965
2	**Czech Republic ('70 recovered)**	**2001**	**1991**	**1980**	**1970**	1961
1	**Germany (FR and DR)**	**2001m**	**1991m**	**1987, 81**	**1970, 71**	1961
2	**Italy ('81 recovered)**	**2001**	**1991**	**1981**	1971	1961
3	**Netherlands ('60 recovered)**	**2001m**			**1971**	**1960**
	Slovenia	**2001**	**1991**	1981		
	Switzerland	**2000**	**1990**	**1980**	**1970**	1960
2	**United Kingdom**	**2001**	**1991**	**1981**	**1971**	**1961**
North America and the Caribbean						
4	**Canada**	**2001**	**1991,** 96	**1981,** 86	**1971,** 76	1961, 6
3	**Dominican Republic**	**2003**	**1993**	**1981**	**1970**	1960
2	**El Salvador**		**1992**		**1971**	1961
5	**Guatemala**	**2003**	**1994**	**1981**	**1973**	**1964**

4	**Honduras**	**2000**		**1988**	**1974**	**1961**
2	**Nicaragua**	**2005**	**1995**		**1971**	1963
5	**Panama**	**2000**	**1990**	**1980**	**1970**	**1960**
4	**Puerto Rico**	**2000**	**1990**	**1980**	**1970**	1960
South America						
3	**Bolivia**	**2001**	**1992**		**1976**	
5	**Paraguay**	**2002**	**1992**	**1982**	**1972**	**1962**
1	**Peru ('81 in recovery)**		**1993**	**1981**	1972	1961
4	**Uruguay ('63 recovered)**		**1996**	**1985**	**1975**	**1963**
Africa						
2	**Egypt**		**1996**	**1986**, 81	1976	1964
	Ethiopia (in progress)	**2007**	**1994**	**1981**		
2	**Ghana** (in progress)	**2000**		**1984***	**1970***	
2	**Guinea, Conakry**		**1996***	**1983***		1960
	Lesotho	**2006**	**1996***	**1986***	1976	1966
1	**Madagascar**		**1993**			
3	**Malawi**		**1997***	**1987***	**1977***	1967
3	**Mali ('76 in recovery)**		**1998**	**1987***	**1976**	
2	**Mauritius**	**2000***	**1990***	**1983**	1972?	1962
2	**Sudan ('73 recovery underway)**		**1993**	**1983**	**1973***	
Asia						
1	**Armenia ('89 lost)**	**2001**		1989	1979	1970
	Bangladesh ('81 to be recovered)	**2001**	**1991**	**1981**	1974	1961
3	***Fiji Islands ('76 in recovery)**	2007	**1996**	**1986**	1976	**1966**
6	**Indonesia**	**2000**	**1990, 95**	**1980, 85**	**1971**	1961
1	***Iraq ('87 destroyed by looting)**	2007	**1997**	1987	1977	1967
4	**Malaysia ('70, '80 recovered)**	**2000**	**1991**	**1980**	**1970**	1960
1	***Mongolia ('89 to be recovered)**	**2000**		1989	1979	1970
3	***Pakistan ('73, '81 to be recovered)**		**1998**	**1981**	**1973**	1961
4	**Thailand (new samples in process)**	**2000**	**1990**	**1980**	**1970**	1960
1	**Turkmenistan ('89 lost)**		**1995**	1989	1979	1970

1. Three goals of the IPUMS-International initiative: preserve, integrate and disseminate census microdata. Census microdata are the individual responses to census questionnaires recorded in computerized form as numeric or alphabetic codes. The data include such mundane characteristics as age, sex, marital condition, relationship to head, migration, education, occupation, etc. Often the datasets include records for families, households and dwellings as well as for individuals. Over the past half century most of the major statistical agencies have prepared census microdata files for analysis by staff and, in many cases, by external researchers. Before the microcomputer revolution of recent years, the computational resources to analyze census microdata were the exclusive preserve of only the official statistical authorities, large universities or well endowed research institutes. Now, with the ever-expanding power of microcomputers, analysis of large census microdata files is readily performed by ordinary researchers and, increasingly, even by their students.

Today, in developed countries, census microdata are widely used by researchers and policy makers, but are relatively little used elsewhere. This gap is about to shrink, thanks, on the one hand, to the IPUMS-International global initiative (https://www.ipums.org/international) led by the University of

Minnesota Population Center and, on the other, to a policy revolution by statistical authorities in the developing world, which are increasingly recognizing census microdata as statistical products to be disseminated along with conventional publications. A good example of this revolution is the dissemination policy of the Central Statistical Agency of the Federal Democratic Republic of Ethiopia, which in 2005 began to distribute a wide variety of microdata products on CD and from its website (http://www.csa.gov.et).

IPUMS-International has three goals: first, to preserve census microdata; second, to integrate anonymized samples; and third, to manage access to sample extracts for researchers and policy analysts free of charge.

Preserving microdata. Data recovery is required for all but the most recent datasets. The recovery of data from old tapes is a challenging undertaking for even the most technically skilled cyber sleuths. The MPC does not recover data. Instead the project pays costs of data recovery, relying on the technical skills and widely recognized talents of the United Nations Demographic Center for Latin America and the Caribbean (CELADE) or, where more convenient, a specialized data recovery firm. Most of the datasets for censuses from the 1960s or 1970s were recovered in this way. For example, in the case of the 1979 census microdata of Kenya, in addition to the five percent national sample held by ACAP, approximately two-thirds of the person records (9,781,690) were recovered by a commercial firm at a cost of less than $1,000. The project's most recent success was the 1977 census of Romania, where 97.2% of the person records were recovered by the same firm.

In the case of Africa, beginning in the 1990s, the African Census Analysis Project (www.acap.upenn.edu) blazed a path, methodically assembling a collection of microdata from some 25 countries for a total of more than 45 censuses (Zuberi 2005). The ACAP repository offers a trove of census data, much of it recovered from old computer tapes. A growing number of researchers and graduate students in residence at the University of Pennsylvania are exploiting these materials, primarily for academic research. An example of the fruits of the initiative is the recently published book entitled "The Demography of South Africa" in collaboration with Statistics South Africa and based on a ten percent sample of the 1996 census (Zuberi, Sibanda and Udjo 2005).

Meanwhile, as of June 2007, IPUMS-International has become the largest repository of census microdata in the world with the official statistical authorities of more than 60 countries, encompassing over six-tenths of the world's population, entrusting a total of 174 censuses to the Minnesota Population Center (Table 1). The successes of ACAP and IPUMS is due in part to the increasing recognition among official statisticians that anonymized census microdata constitute statistical data products.

Integrating Microdata. To make census microdata useful for research they must be thoroughly documented and integrated. International census samples employ differing numeric classification systems and reconciliation of these codes is a major effort. Variables must be easy to use for comparisons across time and space. This requires that we provide the lowest common denominator of detail that is fully comparable. On the other hand, we must retain all meaningful detail in each sample, even when it is unique to a single dataset (Ruggles et. Al. 2003).

For most variables, it is impossible to construct a single uniform classification without losing information. Some samples provide far more detail than others, so the lowest common denominator of all samples inevitably loses important information. Composite coding schemes offer a solution. Similar to those used by the International Labor Organization for occupations and industries, we apply composite coding to each variable to retain all original detail, and at the same time provide

comparable codes across countries and censuses. The first one or two digits of each code provides information available across all samples. The next one or two digits provide additional information available in a broad subset of samples. Finally, trailing digits provide detail only rarely available. Where a concept is not present, a zero place-holder is assigned to that digit.

Consider educational attainment, for example. In the IPUMS-International system, the first digit of this variable with four categories is comparable across all samples: less than primary completed, primary, secondary and university completed. The second digit delineates additional details. The final digit provides additional detail, such as the difference in five and six years of primary schooling completed (Esteve and Sobek 2003). For example for less than primary completed, "211" is the code for "5 years of primary schooling completed", while "212" indicates "6 years completed", "221" indicates lower secondary completed (unspecified track) and "222" technical track completed.

The basic goal of the harmonization efforts is to simplify use of the data while losing no meaningful information. The IPUMS harmonization strategy has proven flexible enough to accommodate the integration of data across broad spans of time (the United States for 1850-2000) and space (Brazil, China, Colombia, France, Kenya, Uganda, the United States, and Vietnam; Sobek et. al. 2002).

Managing Access. Researchers must first be approved before access to any microdata is permitted. Moreover users are never permitted access to the original source files provided by the NSOs. Instead, data are provided in the form of extracts, custom tailored to each researcher's needs. What this means is that there is no distribution of entire datasets by means of compact discs. Since each dataset is custom tailored "collecting" or "boot-legging"datasets is not only illegal, but effectively curtailed.

In 2006, the Economic Commission of Europe published guidelines for Managing Statistical Confidentiality and Microdata Access. An IPUMS-Europe case study, using the specific case of France, is appended to the UNECE report as an example of good practice (please see Appendix). The case study describes how IPUMS manages access to microdata, explains why it is a good practice, identifies the target audience, explains confidentiality measures, specifies the rules and procedures regarding user access, summarizes supporting legislation (for others see http://unstats.un.org/unsd/goodprac/default.asp), and lists strengths and weaknesses as well as bibliographical references. While the IPUMS case study is European in scope, the details are nearly identical for the International project.

To request an extract, the researcher must first sign in by entering the registered password. To create an extract, the user makes a series of selections—country (or countries), census years, samples, variables and sub-populations—by means of point-and-click menus. The researcher selects the country or countries, census years, samples, and variables as well as the form of metadata required for the statistics package to be used (SAS, SPSS, or STATA are supported). The IPUMS-International extract engine also makes it possible to select sub-populations, such as say, females aged 15-19 in the workforce.

Once the selections are complete, there is an opportunity to review or revise before final submission of the request. Then, once submitted, the extract engine registers the request and places it in a data processing queue. When the extract is ready (usually in a matter of minutes), the researcher is notified by email that the data should be retrieved within 72 hours. A link is provided to a password-protected page for downloading the specific extract. Soon an SSL (Secure Sockets Layer) protocol will be implemented at the Minnesota Population Center. After SSL is in place, the data will be encrypted during transmission using a 128-bit encryption standard, matching the level used today by

the banking and other industries where security and confidentiality is essential. The researcher may then securely download the file, decompress it and proceed with the analysis using the supplied integrated metadata consisting of variable names and labels. The metadata are in ASCII format so that a researcher may readily use them with any statistical software.

2. The IPUMS-International Dynamic Metadata (Documentation) System: five clicks to compare any question in any combination of countries and censuses. To use census microdata well, researchers must consult original source documentation to understand census concepts, definitions and nomenclatures. IPUMS has developed a dynamic metadata system to facilitate comparison of the phrasing, in English, of any census question with any combination of countries and years in the database. Five clicks are all that is required. For example, if the researcher wishes to compare the marital status question in the censuses of Mexico with those of the United States and Spain, this is accomplished simply by (1) accessing the web page (www.ipums.org/international), selecting (clicking) (2) "Variables", (3) the countries and census samples (years), (4) the educational attainment variable, and finally (5) "enumeration text". At this point, a screen will appear, displaying the census question, categories, and enumerator instructions—all in English. If the researcher wishes to see the sample frequencies of the codes for any variable, from that variable page (click "4", above), click codes, then "Case-count view" and the frequencies will appear for the selected variable and census samples. For complex variables, click "detailed codes" to see the full range of codes in the IPUMS database for this variable. Images of the original forms and instructions in the official language(s) of each census are also available, by means of 2 clicks from the home page (click "Census Questionnaires", then the specific country, census year and document). Images are not "book-marked", therefore, unlike for the dynamically generated enumeration text pages, scrolling is required to navigate to the item of interest.

Using the dynamic enumeration text pages, the researcher may easily study the displayed documentation to determine if the phrasing of questions in the various censuses is sufficiently alike to permit comparison, given the precise topic of research. For example, in the Ugandan census of 2002, there is no direct distinction between those who never attended primary school and those who attended some years, but did not complete primary instruction. However, this subtle distinction can be determined by examining " years of schooling".

Note that no registration is required to use the IPUMS dynamic enumeration text feature. Unlike access to the microdata, any user may browse IPUMS census documentation without registration. However, dynamic enumeration text is available only for censuses for which the microdata are integrated into the IPUMS database. It is for this reason that when microdata are entrusted to IPUMS, a concerted effort is made to obtain complete documentation—forms, instructions, codebooks, and technical or methodological studies—so that these may be integrated into the metadatabase. For non-English language materials, translators of all the major world languages are contracted by the IPUMS project to provide English translations of source documentation.

For the 2010 census round, some of our NSO partners are considering providing "tagged" census metadata so that they can be transferred to the IPUMS system with minimal additional treatment. We would be delighted to discuss this possibility with delegates of interested NSOs.

3. How census microdata may be used for local planning to attain Millennium Development Goals: schooling, literacy and gender equity in Uganda. The United Nations has an ambitious campaign, *The Millennium Development Goals*, which lays out a total of 8 development objectives to attain by the year 2015. The 191 member countries of the United Nations have endorsed

objectives to eradicate extreme poverty and hunger, achieve universal primary education, promote gender equality and female autonomy, reduce infant mortality, improve maternal health, combat HIV/AIDS, malaria and other diseases, guarantee the sustainability of the environment, and foment a world association for development.

For each objective, the United Nations has developed a battery of indicators to evaluate the situation and to measure improvements in each region and country of the world. Nevertheless, region or nation is not always the most appropriate scale for this type of analysis, because, often, statistics for an entire country are not representative of the situation in small areas or localities, above all in those countries where great inequalities are observed at the local level. For this reason, analysis at the local level can identify the most disadvantaged areas to organize a better distribution of assistance and resources devoted to solving the problems.

Confronted with this challenge, local statistics are called upon to play a more important role in measuring results. Population censuses, and by extension the corresponding microdata, are also a most important source for this type of analysis because they guarantee a more or less homogeneous treatment and complete territorial coverage. Moreover, where more than one census is available, they provide chronological depth so that achievements, or the lack thereof, may be gauged over time.

To illustrate the use of census microdata, we analyze high precision samples for the years 1989-2002 from the population censuses of Uganda available from the Minnesota Population Center at www.ipums.org/international. The original microdata come from the Ugandan Bureau of Statistics. We address the second and third objectives of the Millennium Development Goals (MDGs): to achieve universal primary education and promote gender equity. The analysis focuses on districts, of which, in the case of Uganda, there are 56 identified in the samples. For each district the index proposed by the United Nations is computed. The results are show that the principal deficiencies are confined to a few districts with the worst conditions.

Is primary education universal in Uganda? To respond to this question we much use three distinct indicators, following UN recommendations. First is the net rate of primary schooling. Calculation of this indicator requires three variables: age, school attendance, and level of education attained. All these variables are available in the microdata for the 1991 and 2002 censuses of Uganda. For purposes of international comparison, we use the UNICEF definition of primary school (6 years) instead of the Ugandan standard of seven (for an explanation, see www.ipums.org/international "Variables", "EDATTAIN"; note that the IPUMS variable "YRSCHOOL" facilitates comparing any definition compatible with the coding of the original source data). The rate is the result of obtaining the percentage of persons attending primary school divided by the total children of primary school age. To compute this indicator, we have taken into account children aged 6-11 years old (1.9 million in 2002). We find that, in 2002, 84% of boys and 81% of girls of primary school age are declared as attending school, up from 69 and 56%, respectively in 1991. This is truly an extraordinary transformation of primary educational opportunities in Uganda, in barely a decade. While the total number of children aged 6-11 years increased by 50%, the absolute number of those not attending school dropped by fully one-third. In 53 of 56 districts enrollment rates are greater than 67%. In fact 80% is the norm for most districts. Enrollment rates of less than 25% are found in only 3 districts (Kotido, Moroto, and Nakapiripiriti). Unfortunately, 332 thousand primary aged children still were not attending school in 2002. Because most districts show a deficit of 10-15 percent, much effort will be required across all districts to attain the goal of universal primary education.

According to the UNICEF standard, six years of schooling is defined as completing primary education. Since longitudinal data, which would permit tracing the evolution of educational attainment of each cohort of students, are not available, we have opted to compute, as an approximation, the percentage of children between age 13 and 15, which had completed their primary studies. Age and level of educational attainment are required to compute this indicator.

The national average from the 2002 census microdata indicates that 33.9% of Ugandan children between the ages of 13 and 15 years have completed their primary schooling, an increase of exactly 10 points in eleven years. Girls enjoyed a three point advantage in 2002, up from 2 points in 1991. These figures are significantly distant from the 100% objective. Nevertheless, this figure is a lagging indicator, since it reflects educational accomplishments (or lack thereof) of some 5-10 years before. Five to ten years hence we would expect these figures to increase sharply. Districts that score poorly on this goal are precisely those where large fractions of children are not attending school at all, and vice-versa. The list of poorly performing districts, with completion rates of less than 7%, is identical to that for primary school attendance. Religious differentials are striking. Of the quarter million Muslims, fully 38% of boys and 46% of girls indicate six years of schooling completed.

Literacy was inquired of in both Ugandan censuses with a slightly more detailed question in the more recent census (Can [Name] read and write a simple sentence in any language?) than in the earlier one (Can person read or write?). The literacy rate for the Ugandan population as a whole aged nine years rose substantially from 53 to 66%, yet this figure remains well short of the Millennium Goal. If we choose 75% as the threshold for progress toward the goal only 8 of 56 districts rise above the threshold: Kalangala, Kampala, Masaka, Mpigi, Mukono, Wakiso, Jinja, and Rukungiri. Muslims and Penecostals are closest to attaining the threshold with 70% of adherents indicated as literate in each case. For Muslims, of almost two million adherents, 76% of males and 66% of females are literate. Penecostals account for 785,000 adherents with 77% of males literate compared with 65% of females. Among ethnic groups the Baganda, numbering almost 3 million and twice as numerous as the next largest group, stand out as the most literate in Uganda (86 and 83% for males and females, respectively).

Gender equity is considered here only partially, since we treat only those aspects related to educational attendance, attainment and literacy, taking as our point of departure the statistics calculated above. In recent decades Uganda has made substantial progress in providing gender equity in schooling. Female rates of primary school attendance slightly exceed male rates in 37 of 69 districts. Discrimination against females in basic access to education is substantial in only five districts (Garissa, Kilifi, Malindi, Marsabit, and Moyale) where male attendance rates exceed those of females by nine or more percentage points. Differences in the lagging indicators of primary school completion and literacy are more noticeable. The 2012 census will reveal the degree to which these gender inequities persist. Meanwhile the more serious problem seems to be general access to education rather than gender equity.

4. Conclusions. Census microdata are exceedingly useful for analyzing populations. Because they are microdata they register the characteristics of individuals and thus can be studied by taking into account any or all of the characteristics present in the record. Because they come from a census, this is a source without paragon for demographic and social analysis due to its high density, complete national coverage, and near simultaneous execution. Moreover, if the microdata are integrated with censuses from several decades and different countries, comparative analysis in time and space opens additional avenues for research. In sum, integrated census microdata are destined to play an

important role in social science research and policy making, as has been demonstrated here with the example of the Millennium Development Goals. Without doubt, the use of census microdata will have a significant, positive impact on the understanding of the social and demographic dynamics of individuals, families, and nations.

The IPUMS-International initiative is conscious of this potential, and it is for this reason that the National Science Foundation is providing sustained funding to develop a global collaboratory with national statistical authorities, universities, and research institutes. Institutions and researchers interested in working on this initiative to add more samples for more countries are invited to contact the authors of this paper. Researchers interested in using the microdata are invited to apply for access and use the microdata as research needs require.

References.

Esteve, A. and Sobek, M.. (2003). Challenges and Methods of International Census Harmonization. Historical Methods 36: 66-79.

Ethiopia, Democratic Federal Republic. Central Statistical Agency. (2004).“Directive No. 1/2004. Directive issues to establish procedures for accessing raw data to users,”. http://www.csa.gov.et/text_files/directives.htm

McCaa, R. and Ruggles, S. (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.

Ruggles, S., Sobek, M., McCaa, R., King, M. and Levison, D. (2003). IPUMS-International. Historical Methods 36: 60-65.

Ruggles, S, and Sobek, M., et. al. (1997). Integrated Public Use Microdata Series: Version 2.0. Minneapolis: Historical Census Projects, University of Minnesota.

Minnesota Population Center. (2007). Integrated Public Use Microdata Series-International: Version 3.0. Minneapolis: University of Minnesota.

Statistics South Africa (2003). 2001 Census of Population and Housing: 10% Sample of unit records (Version 1).

Uganda Bureau of Statistics (2004). 2002 Census of Population and Housing. 10% Sample of unit records.

Zuberi, T. (2005). “Building regional data archives: the African Census Analysis Project (ACAP),” IUSSP XXV International Population Conference, Tours France, July 18.

Appendix. Case Study for **Managing Statistical Confidentiality and Microdata Access: Principles and Guidelines of Good Practice**: http://www.unece.org/stats/documents/tfcm/1.e.pdf
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 Demogràphics, Autonomous University of Barcelona) with France, as a specific example.

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 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). Five other European countries have endorsed the project, but have not yet provided data: Bulgaria, Germany, Italy, Slovenia, and Turkey. Beginning in 2008, 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 exclusively for statistical purposes associated with teaching, research, and publishing;
3) use for administrative, 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 principal sanction for misuse is recall of data and an embargo against use by the individual and the individual's institution. In addition, the sanctions clause of the MOU threatens additional sanctions to assure compliance:

"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. Offers security against loss of source microdata. Raw data files entrusted to the project are encrypted and stored in a secure data repository. Copies of these files are made available only to the National Statistical Office-owner, and are never re-distributed to others.
b. Fosters maximum uniformity of approach and facilitates greater access to microdata by the research community.
c. Improves on arrangements for providing access to microdata to the greater satisfaction of both the National Statistical Offices and the research community.
d. National Statistical Offices cede census microdata files to the University of Minnesota. The data are anonymized and then integrated. Much new integrated metadata are written and stored in a database accessible to all at no cost via the internet. Integrated microdata are available 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.
e. 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.
f. 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.
g. 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 recall of data and sanctions against both the individual and his/her institution.
h. 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 to the extent permitted by national legislation.

i. Microdata are encrypted during transmission using 128-bit SSL (Secure Sockets Layer) encryption standards used by the financial industry.

j. 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.

k. 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.

l. 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.

m. 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. Users do not have access to original source files supplied by the National Statistical Office. Instead researchers access integrated microdata with codes and documentation which not only may differ from the original source but also may contain errors introduced in the integration process.

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.