IPUMS-International High Precision Population Census Microdata Samples: Balancing the Privacy-Quality Tradeoff by Means of Restricted Access Extracts

Robert McCaa

Minnesota Population Center, 50 Willey Hall
Minneapolis MN 55455 USA
contact: rmccaa@umn.edu

Abstract. A breakthrough in the tradeoff between privacy and data quality has been achieved for restricted access to anonymized population census microdata samples. The IPUMS-International website, as of June 2007, offers to approved researchers integrated microdata for 26 countries (80 censuses), totaling more than 200 million person records. Samples for 60-80 additional censuses will be integrated over the next four years by the global collaboratory led by the Minnesota Population Center. Major funding is provided by the United States National Science Foundation and the National Institutes of Health. The statistical authorities of more than 70 countries have already entrusted microdata to the project under a uniform memorandum of understanding which permits researchers to obtain custom extracts without charge and to analyze the microdata using their own hardware and software. This paper describes the disclosure control methods used by the IPUMS initiative to protect privacy and to enhance quality in providing access to high precision census microdata samples.

Keywords: Census microdata samples, data privacy, data quality, IPUMS-International

1 Introduction

In 1983, the legendary Charles M. Cawley offered the alumni association of his alma mater, Georgetown University, a deal. In exchange for its endorsement and a list of members, his fledgling credit card company, MBNA, would pay a percentage of revenues to the association. The offer was accepted and MBNA—by extending the affinity credit card offer to organizations with responsible, affluent members (from the Association of Trial Lawyers of America to the Sierra Club)—quickly established itself as the fastest growing, most profitable credit card company in the United States. Cawley became a billionaire. Now every successful credit card company in the world markets affinity cards.

The IPUMS project seeks neither profits nor popularity. Ours is a wholly academic initiative, but we target an affinity group, a "restricted class of individuals" [1] consisting of academic and policy researchers, who have great need to use population census microdata, but pose a vanishingly small risk of misuse.

2 What is IPUMS-International and why is it a "good practice"?

IPUMS-International is a collaboratory of universities and national statistical offices led by the University of Minnesota Population Center. The project, which began in 1999 with funding by the National Science Foundation of the United States, has three goals:

1) Preserve census documentation and microdata
2) Integrate anonymized census samples using uniform concepts and coding designs
3) Disseminate census sample extracts and documentation to bona fide researchers at no cost

Table 1 provides an inventory of microdata entrusted to the IPUMS project, by country and census year, indicating the sample size and design of 80 anonymized samples already integrated into IPUMS database, plus 109 other census datasets at various stages of integration. It should be noted that in many instances the statistical authorities have provided 100% microdata for preservation and the construction, by the IPUMS project, of anonymized samples according to uniform methods. Others have recovered data for historical censuses and drawn new, high precision samples for the database. Still others are in the process of recovering microdata and documentation for long-forgotten censuses.

Where much disclosure control research on the privacy-quality tradeoff is focused on either "public access" at one extreme or "safe-harbor" at the other [2], the IPUMS-International initiative adopts a third way, the "trusted user" approach [3]. Access is denied to approximately one-third of those who complete the electronic application form. Five years after dissemination began in May 2002, fewer than two thousand researchers have been granted access to IPUMS-International census microdata.

We restrict access to researchers who have a defined need to use the data and who not only agree to abide by the rigorous conditions of use license but also bind their institutions as enforcing agents. With, on the one hand, the assistance of our statistical agency partners, as stipulated in the project memorandum of understanding, and, on the other, the conditions of use license, misuse will lead to punishment not only for the individual but also for the individual's institution. Indeed, in contrast to the record of commercial companies and government agencies, where there are frequent stories of misuse of microdata for disclosing information about individuals, there is not a single, specific allegation of misuse of population census microdata in more than four decades of use by academic researchers. By rigorously policing access, we expect to extend this unblemished record of responsible scholarly use. IPUMS pays a license fee to the National Statistical Office in order that the samples may be disseminated at no cost to researchers.

In 2007, the IPUMS-International initiative became the only academic project cited as "good practice" by the UNECE joint task force on Managing Statistical Confidentiality and Microdata Access: Principles and Guidelines of Good Practice [4]. As the case study explains, the IPUMS constitutes good practice because it provides a high level of trust and transparency for both researchers and National Statistical Offices. Conditions of use are carefully defined and sanctions for misuse clearly spelled out. Usage is enforceable by a strong set of administrative and legal mechanisms. The data are confidentialized by powerful technical measures, previously not considered by National Statistical Offices. The case study is appended to this report.

It must be emphasized that the IPUMS project invites cooperating National Statistical Office to draw fresh, high precision samples for their entire sweep censuses for which microdata are extant or recoverable. Many Offices have responded positively to the invitation, including FSO-Germany, INSEE-France, ONS-UK, ISTAT-Italy, INEGI-Mexico, IBGE-Brazil, CAPMAS-Egypt, BBS-Bangladesh, NSO-Philippines, and many others.

National Statistical Offices that disseminate samples often limit documentation to a simple codebook. Inadequate documentation leads to poor quality research. The IPUMS project promotes high quality research by providing not only codebooks, but also copies of the census forms and instructions to enumerators as well as extensive metadata describing the samples, variable definitions, comparability discussions, etc. Documentation is available in the official language as well as English translations, commissioned by the project where none exists. All the documentation is available in a dynamic metadata system which can be used to compare the actual phrasing of census questions and instructions, for any combination of countries and census years.

3 The Case for High Precision Samples: The USA Experience

In recent years, scholars working with United States census microdata have come to rely on high-precision samples. Beginning with the 1980 census, the Census Bureau has released five-percent samples as well as the one-percent samples. The five-percent samples for the United States in 1980, 1990, and 2000 include between 12 million and 14 million individuals in each year.

The Census Bureau anticipated that the 1980 five-percent sample would be used mainly for state and local policy analysis; at the time the sample was created, it was prohibitively expensive for most researchers to process the entire set of five-percent data. By the end of the 1980s, however, data processing costs had declined dramatically and were no longer a critical constraint for researchers at major institutions. Social scientists soon developed research strategies that capitalized on the availability of very large census microdata files. Swicegood et al. [5] published the first article in *Demography* that used a five-percent national sample, an analysis of language use and fertility in the Mexican-origin population. Later that year, Odland and Ellis [6] published a second *Demography* article using the large 1980 file, a study of household size and regional outmigration rates between 1975 and 1980.

From that time on, the use of high-precision census microdata files expanded rapidly. The cost of computing declined dramatically during the first half of the 1990s with the advent of inexpensive UNIX workstations. Moreover, during the past several years the performance of Windows-based desktop computers has improved to the point that a machine costing less than $1,000 is now easily capable of processing the five-percent samples of 1980, 1990 and 2000. Since 1996, the on-line data dissemination systems developed at Minnesota and elsewhere have provided easy access to large microdata extracts. Accordingly, the largest census microdata files—once available to few researchers at great expense—are now accessible, at no cost, to virtually all social scientists and policy analysts worldwide.

Increasingly, studies that use census microdata from 1980, 1990 or 2000 have turned to the five-percent files. Since 1990, 81 percent of *Demography* articles based on recent census microdata have used the high-precision samples.[1] Most of these analyses depend on information for small population subgroups, ranging from same-sex couples to the grandchildren of immigrants. In many instances, the large samples permit the use of innovative methods; to take just one example, these files have allowed demographers to carry out multi-level contextual analyses by making it feasible to assess the characteristics of small geographic areas.

The five-percent samples of the 1980, 1990 and 2000 censuses have now become the most widely used data source in the pages of *Demography*, as we learned from a analysis of the journal's pages in 2002. At that time, even though the United States had abundant high-quality survey data and the most recent census samples were over a decade old, high-precision census microdata files were used by a quarter of the articles on the United States that appeared in *Demography* in 2000 and 2001. In that period, the large samples were used twice as often as the next most popular data source. Clearly, the high-precision samples of the 1980 and 1990 censuses had become an indispensable component of American social science infrastructure. In 2003, with the addition of a five percent sample from the 2000 census, use skyrocketed.

It is impossible to determine an optimal size for a general-purpose sample. The number of cases needed to analyze a population subgroup depends on desired precision, type of subgroup, type of analysis, and population heterogeneity. If high precision estimates are required, many thousands of cases of the subgroup of interest may be necessary. Frequently, the relevant individuals for analysis are a small subset of the sample population. Multilevel analyses of the effects of local context on individual behavior are especially demanding since they often require data tabulated for small geographic units. The experience of the U.S. demonstrates that very large census microdata samples are among the most powerful tools available for economic and demographic analysis. As such samples become available for other countries around the world, they are becoming key components of social science and policy infrastructure.

4 The IPUMS Approach: High Precision Samples with Implicit Stratification

An important technique used to protect confidentiality of census microdata is to draw a high precision sample from all the census microdata records and then, in addition to the disclosure controls discussed below in sections 4 and 5, suppress from the sampled records all identifying information (names, addresses, and low-level geographical details). High precision samples preserve the ability to work with a large amount of microdata making it harder to identify any one person in the sample data file. In drawing high precision samples it is also important to think about efficient methods. By using stratification to draw a high precision sample, gains in efficiency are possible [7], [8]. To the extent the strata used to draw a high precision sample are associated with the variables of interest (e.g., orphanhood, poverty, unemployment, etc.), the resulting estimates of these variables will have lower standard errors than what would have resulted had a simple random sample of records been drawn from the complete census data [7], [8].

One of the most important stratifying variables in survey research and in drawing high precision census microdata samples is geography. Geography is related to a great number of variables researchers are interested in studying and therefore increases the efficiency of stratified samples. Many of the IPUMS-International samples capitalize on *implicit* geographic stratification. The raw census files used to create IPUMS samples are typically geographically organized within districts. Systematic random samples of the censuses capitalize on this low-level geographic sorting. By ensuring a representative geographic distribution of sampled cases, they are equivalent to extremely fine geographic stratification with proportional weighting. Since many economic and demographic characteristics are highly correlated with geographic location, this implicit stratification yields substantially greater precision than would a simple random sample of households. As part of the IPUMS project, we are developing stratification variables that allow researchers to make reliable variance estimates from implicitly stratified samples.

Almost all the statistical agency partners of the IPUMS project have endorsed the use of implicitly stratified samples of households (see Table 1, "sample design" column). Twenty-six countries (identified by "*" in Table 1) have provided complete sets of census microdata to facilitate the drawing of implicitly stratified samples by the project. In Europe, almost all the statistical agencies have drawn new samples using IPUMS specifications. IPUMS sample densities, as can be seen in Table 1, typically range between 5 and 10%. Lower densities are provided by countries where privacy matters are a greater issue than quality (Netherlands, United Kingdom) or, as in the case of 1960 round of censuses, where only low precision samples survive.

[1] This percentage excludes eight articles that did not specify sample precision.

5 IPUMS-International Access Disclosure Controls

Access to the IPUMS-International database is governed, on the one hand, by the letter of understanding endorsed by the University and the National Statistical Authority, and, on the other by the license agreement between the University, the researcher, and the researcher's institution. Both are subject to amendment and enhancement as new methods are suggested. The letter of understanding grants the right to the university to disseminate microdata extracts electronically for teaching and research purposes via the project webpage: https://www.ipums.org/international, according to the authorization procedures stated in the agreement. Data may not be used for commercial purposes. Strict confidentiality of persons, households and other entities must be maintained. Alleging that a person or other entity has been identified is prohibited. The University is charged with assuring that users will guard against access to the microdata by unauthorized individuals.

The fact that IPUMS-International distributes microdata electronically as custom extracts, tailored as to country(ies), census year(s), subpopulation(s), and variables, according to the individual needs of the researcher, provides additional incentives for jealously guarding extracts. Since complete datasets are not distributed on CD or other medium, the inclination to share data with unauthorized individuals is greatly reduced—even completely eliminated.

The electronic application form is designed to ascertain the bona fides of the applicant as well as the appropriateness of the microdata for the proposed research. A stern warning is issued against fraudulent applications, and checks are implemented to verify the identity and affiliations of the applicant (see the project home page "Apply for Access"). To confirm that the researcher understands the sensitivity of guarding the privacy of individuals, the application requests the name of the Human Subjects Protections Committee, Institutional Review Board, or similar office at the applicant's institution. A critical consideration in determining access is the proposed research. The statement must identify the data to be used and the purpose. Many applicants are denied access for failing to demonstrate that microdata are needed to address the proposed research or instructional plan. Finally the researcher must agree to seven restrictions on use: no redistribution, scholarly use only, prohibition on commercial use, strict rules of confidentiality, data security, appropriate citation, and notification of errors in the data. Approval is granted for a period of one year and may be renewed. Access to the microdata is password controlled. Remote data access is not offered. While this method might allow access to higher density, virgin microdata, our memorandum of understanding with the national statistical agencies does not authorize this form of access.

6 Technical Disclosure Controls

Where the statistical agency entrusts the anonymization procedures to the IPUMS project, we impose additional technical privacy protections. Technical controls are implemented on a subjective, ad-hoc basis as negotiated with each country for each census. Contemporary microdata, say from a census taken less than ten years ago, require more technical disclosure controls than older, historical data.

The most important technical control is the suppression of records by subsampling. All the values in the records outside the sample are suppressed. Second, is the suppression of names and geographical detail, such as place of birth or residence. Each statistical authority balances the trade-off by instructing the IPUMS project as to the minimum threshold for identifiable geographical units for the most recent census. In the case of many African and Latin American countries, the threshold is commonly set at 20,000 inhabitants in the latest census. Others place it as high as 100,000 (United States) or in the most extreme case (Netherlands) all administrative geography is suppressed. We are gratified that in some cases our statistical agency partners have reconsidered earlier decisions, offering higher precision samples (Mexico 1990 increased from one to ten percent) and greater detail. In the case of Colombia, the geographical threshold, initially set at 100,000, was reduced to 20,000 after Colombian geographers vigorously registered their dissatisfaction. The Colombian statistical agency not only reduced the threshold, but also harmonized the identifiers so that all the census microdata samples for Colombia could be disseminated with a single set of geographical codes.

Additional protection is provided by randomly ordering the records and swapping the geographical identifiers of an undisclosed number of households. This means that no one can state with certainty that an individual or household has been identified.

In consultation with the national statistical office, some variables may be top-coded, others may be subjected to global recoding, deletion of digits for hierarchical variables (occupation, industry, geography), or the suppression of a variable entirely. Decisions are made in consultation with the corresponding national statistical authority. Sensitive variables, if any, may be suppressed entirely at the request of the statistical agency. Weight variables are usually not an issue because most of the samples are implicitly stratified with a single weight. We do not resort to either microaggregation or Post Randomization (PRAM) methods.

7 Countering Fear, Hysteria and Paranoia with Reason

Privacy rights and statistical confidentiality of data are severely threatened by government, commercial firms, and individuals—but the threat to population census microdata is virtually nil. Fear, hysteria and paranoia are incited among official statisticians by the widespread circulation of a "pizza commercial" developed by an American civil liberties advocacy group [9] and advertisements offering private details of individuals and entities for a price. What is striking is that none involve population census microdata. Indeed, there is no market—black, grey, gold or otherwise—for anonymized census microdata samples for the purpose of identifying individuals or linking to other data sources. Even in the United States, at a moment of shocking violations of individual rights by government agencies, there is not one allegation of access to census microdata by the Homeland Security Agency or other US government agencies. The reason is obvious. Population census microdata samples, per se, do not contain sensitive or valuable political or commercial information, and without personal identifiers, statistical linkage is useless due to the high proportion of false positives [10]. In the case of Colombia, the first country to join the IPUMS-International intiative, hundreds of Colombian researchers are now using IPUMS integrated census samples to the great satisfaction of the Colombian National Statistical Authority, DANE. Recently the contract with DANE was extended so that samples for the 2005 census could be entrusted to the IPUMS project for integration and dissemination.

8 Conclusion

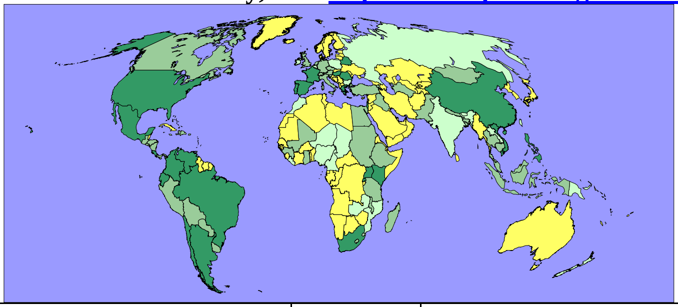
The goal of IPUMS is to restore balance to the privacy-quality tradeoff by providing high precision, anonymized samples to a restricted class of researchers. In the IPUMS datasets identification is impossible for the vast majority of persons and positive identification is always impossible. Given the wealth of information readily available from private sources in most countries, it would be foolhardy to turn to census microdata to attempt to uncover imprecise and outdated information about a particular individual. We invite academics who need census microdata for research purposes to examine the offerings at the IPUMS website. We invite National Statistical Offices that have not yet joined the IPUMS global collaboratory to consider the benefits to be gained by entrusting high quality samples to the project for integration and dissemination.

Acknowledgements. Funded in part by the National Science Foundation of the United States, Grant No. SES0433654.

References

1. Willenborg, L., de Waal, T.: Elements of Disclosure Control. New York: Springer-Verlag (2001)
2. Willenborg, L., de Waal, T.: Statistical Disclosure Control in Practice. New York: Springer-Verlag (1996)
3. McCaa, R., Esteve, A.: IPUMS-Europe: Confidentiality measures for licensing and disseminating restricted-access census microdata extracts to academic users. In Monographs of official statistics: Work session on statistical data confidentiality. Luxembourg: Office for Official Publications of the European Communities, (2006) 37-46.
4. United Nations Economic Commission for Europe, Conference of European Statisticians. Managing Statistical Confidentiality & Microdata Access: Principles and Guidelines of Good Practice. New York and Geneva, 2007: http://www.unece.org/stats/documents/tfcm.htm (see Annex 1.23).
5. Swicegood, G., Bean, F.D., Stephen, E.H., Opitzm, W.: Language Usage and Fertility in the Mexican-Origin Population of the United States. Demography. 25 (1988) 17–33
6. Odland, J., Ellis, M.: Household Organization and the Interregional Variation of Out-migration Rates. Demography. 25 (1988) 567-579
7. Kish, L.: Weighting for Unequal P_i. Journal of Official Statistics. 8 (1992) 183-200
8. Kish, L.: Survey Sampling, Wiley Classics Library Edition. New York: Wiley and Sons (1995)
9. American Civil Liberties Union (ACLU). Surveillance Campaign. (2005) Available online at http://www.aclu.org/pizza/
10. Dale, A., Elliot, M.: Proposals for 2001 SARS: An assessment of disclosure risk. Journal of the Royal Statistical Society. Series A. 164, part 3 (2001) 427-447

Table 1. IPUMS-International: 189 microdatasets entrusted by country, census, sample size and design
For current data availability, see: https://www.ipums.org/international



Sample size			Country	Sample design	Census decade				
10%	~5%	<=4%			2000s	1990s	1980s	1970s	1960s
Released 2002-2007 (26 countries, 86 censuses)									
4			**Argentina**	IPUMS	**2001**	**1991**	**1980**	**1970**	1960
1			**Belarus**	IPUMS		**1999**	1989	1979	1970
5			**Brazil**	IPUMS	**2001**	**1991**	**1980**	**1970**	**1960**
1			**Cambodia**	IPUMS		**1998**			1962
4		1	***Chile**	IPUMS	**2002**	**1992**	**1982**	**1970**	**1960**
		2	**China (‘90 in preparation)**		**2000**	**1990**	**1982**		1964
3		2	***Colombia (’05 in prep.)**	IPUMS		**1993**	**1985**	**1973**	**1964**
3	1		***Costa Rica**	IPUMS	**2000**		**1984**	**1973**	**1963**
4		1	***Ecuador**	IPUMS	**2001**	**1990**	**1982**	**1974**	**1962**
	6		**France (’99 in prep.)**	IPUMS	**1999**	**1990**	**1982**	**1975**	**1968, 2**
4			**Greece**	IPUMS	**2001**	**1991**	**1981**	**1971**	1961
	4		**Hungary**	IPUMS	**2001**	**1990**	**1980**	**1970**	
3			**Israel**	IPUMS		**1995**	**1983**	**1972**	1961,7
	3		**Kenya (’79 in prep.)**	IPUMS	**1999**	**1989**	**1979**	**1969**	
2		2	**Mexico (’80 in recovery)**	IPUMS	**2000**	**1990**	1980	**1970**	**1960**
1			**Palestinian Authority**	IPUMS		**1997**			
3			***Philippines (’70 in prep.)**	IPUMS	**2000**	**1990**	**1980**	**1970**	**1960**
	3		**Portugal**	IPUMS	**2001**	**1991**	**1981**	1970	1960
3			**Romania (’77 in prep.)**	IPUMS	**2001**	**1992**		**1977**	1965
2			***Rwanda**	IPUMS	**2002**	**1991**			
2			**South Africa**		**2001**	**1996**-1	1985-0	1970	1960
	3		**Spain**	IPUMS	**2001**	**1991**	**1981**	1970	1960
2			***Uganda**	IPUMS	**2002**	**1991**	1980		1969
	5		**United States**		**2000**	**1990**	**1980**	**1970**	**1960**
4			***Venezuela (’01 in prep.)**	IPUMS	**2001**	**1990**	**1981**	**1971**	1961
	2		**Vietnam**	IPUMS		**1999**	**1989**	1979	
Europe (total: 20 countries, 71 censuses)									
4			**Austria**	IPUMS	**2001**	**1991**	**1981**	**1971**	1961
			Bulgaria (in process)		**2001**	**1992**	**1985**	1975	1965
	2		**Czech Republic**	IPUMS	**2001**	**1991**	**1980**	**1970**	1961
1			**Germany** (in process)	IPUMS	**2001m**	**1991m**	**1981-7**	**1970-1**	1961
			Ireland (negotiating)		**2001**	**1991**	**1981**	**1971**	
			Italy (in process)	IPUMS	**2001**	**1991**	1981	1971	1961
		3	**Netherlands**		**2001m**			**1971**	**1960**
			Poland (negotiating)		**2001**		**1988**	**1970-8**	1960

			Russia (negotiating)		**2002**		**1989**	1979	1970
			Slovenia	IPUMS	**2001**	**1991**	**1981**		
			Switzerland (negotiating)	IPUMS	**2000**	**1990**	**1980**	**1970**	1960
			Turkey (in process)	IPUMS	**2000**	**1990**	1980-**5**	1970-5	1960, 5
		2	**United Kingdom**		**2001**	**1991**	**1981**	**1971**	**1961**
North America and the Caribbean (12 countries, 45 censuses)									
		4	**Canada**		**2001**	**1991**-6	**1981**-6	**1971**-6	1961, 6
1	1	2	***Dominican Republic**	IPUMS	**2003**	1993	**1981**	**1970**	**1960**
1			***El Salvador**	IPUMS		**1992**		1971	1961
2		3	***Guatemala**	IPUMS	**2002**	**1994**	**1981**	**1973**	**1964**
3		1	***Honduras**	IPUMS	**2000**		**1988**	**1974**	**1961**
2		1	***Nicaragua**	IPUMS	**2005**	**1995**		**1971**	1963
5			***Panama**	IPUMS	**2000**	**1990**	**1980**	**1970**	**1960**
	4		**Puerto Rico**		**2000**	**1990**	**1980**	**1970**	1960
2			**Saint Lucia**	IPUMS	**2001**	**1991**	**1980**	1970	1960
South America (10 countries, 39 censuses)									
3			***Bolivia**	IPUMS	**2001**	**1992**		**1976**	
4		1	***Paraguay**	IPUMS	**2002**	**1992**	**1982**	**1972**	**1962**
1			***Peru**	IPUMS		**1993**	1981?	1972	1961
4			***Uruguay**	IPUMS		**1996**	**1985**	**1975**	**1963**
Africa (16 countries, 28 censuses)									
2			***Egypt**	IPUMS	2006	**1996**	**1986**	1976	1964
1			***Ethiopia**	IPUMS	2007	**1994**	**1984**		
2			***Ghana**	IPUMS	**2000**		**1984**	1970	
2			***Guinea, Conakry**	IPUMS		**1996**	**1983**		1960
			Lesotho (in process)		**2006**	**1996**	**1986**	**1976**	1966
1			***Madagascar**	IPUMS		**1993**			
2			***Malawi**	IPUMS		**1997**	**1987**	**1977**	1967
3			***Mali**	IPUMS		**1998**	**1987**	**1976**	
2			***Mauritius**	IPUMS	**2000**	**1990**	1983	1972	1962
1			***Sierra Leone**	IPUMS	**2004**		1985?	1974	1963
3			***Sudan**	IPUMS	**2007**	**1993**	**1983**	**1973**	
1			**Tanzania**	IPUMS	**2002**		**1988**	**1978**	1967
Asia and Oceania (17 countries, 39 censuses)									
1			**Armenia**	IPUMS	**2001**		1989	1979	1970
1			***Bangladesh** (in process)	IPUMS	**2001**	**1991**	**1981**	1974	1961
3			***Fiji Islands** ('76 in prep.)	IPUMS		**1996**	**1986**	**1976**	**1966**
7			**Indonesia**	IPUMS	**2000**	**1990**	**1980**	**1971**	1961
1			***Iraq**	IPUMS		**1997**	1987	1977	1967
		4	**Malaysia**		**2000**	**1991**	**1980**	**1970**	1960
1			***Mongolia** ('89 in prep.)	IPUMS	**2000**		**1989**	1979	1970
1			**Nepal**		**2001**	1991?	1981?	1971	1961
3			***Pakistan**	IPUMS		**1998**	**1981**	**1973**	1961
		4	**Thailand** (new samples in prep.)		**2000**	**1990**	**1980**	**1970**	1960
1			**Turkmenistan**	IPUMS		**1995**	1989	1979	1970
Note: **bold country** = Agreement signed between University of Minnesota and National Statistical Authority Year = census; **Bold year** = microdata survive; * = 100% long form microdata entrusted to IPUMS; m=microcensus IPUMS systematic sample design for private households: every nth household stratified by enumeration district.									

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 Demogràphics, 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 micrdoata 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.

McCaa, Robert and Steven Ruggles. 2002. "The Census in global perspective and the coming microdata revolution," Scandinavian Population Studies, 13:7-30.

McCaa, Robert and Wendy L. Thomas. 2003. "Archiving Census Documentation and Microdata: Preserving Memory, Increasing Stakeholders", Notas de Población XXIX(75):303-320

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, Steven Ruggles, Michael Davern, Tami Swenson, Krishna Mohan Palipudi. 2006. "IPUMS-International High Precision Population Census Microdata Samples: Balancing the Privacy-Quality Tradeoff by Means of Restricted Access Extracts," Privacy in Statistical Databases. Berlin: Springer, pp. 375-382.

McCaa, Robert, Steven Ruggles, Matt Sobek, and Albert Esteve. 2006. Using integrated census microdata for evidence-based policy making: the IPUMS-International global initiative, African Statistical Journal, 2(May):83-100.

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