UNITED
NATIONS

E

Economic and Social
Council

Distr.
GENERAL

ECE/CES/GE.41/2009/23
14 October 2009

Original: ENGLISH

ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Group of Experts on Population and Housing Censuses

Twelfth Meeting
Geneva, 28-30 October 2009
Item 5 of the provisional agenda

CENSUS OUTPUT TO MEET USERS NEEDS

Entrusting census microdata and metadata for timely integration and dissemination via the IPUMS-EurAsia and IECM initiatives, 2010-2014*

Note by the Minnesota Population Center, Minneapolis, MN USA and Centre d'Estudis Demogràfics, Autonomous University of Barcelona

"Without question IPUMS-International meets the four Core Principles outlined in CES [Conference of European Statisticians] (2007). It is cited in CES (2007) as a Case Study of good practice. This review confirms its status as good practice for Data Repositories. Indeed it is likely to provide the best practice for a Data Repository for international statistical data."
—Dennis Trewin (2007)
www.hist.umn.edu/~rmccaa/ipums-global/trewin_report_2007.pdf

* Research for this paper was funded in part by the National Institutes of Health of the United States, grant HD047283 European and Asian census microdata harmonization project (IPUMS-EurAsia).and Harmonizing Integrated European Census Microdata (HIECM), funded by the European Union, Research Infrastructures Action, FP6- 026033.

I. SUMMARY

1. Integrated, anonymized census microdata and metadata for 44 countries are presently being disseminated from the IPUMS-International web-site by the Minnesota Population Centre (MPC). A subset of 12 European countries, harmonized to European standards, is accessible through the Integrated European Census Microdata (IECM) portal at the Universidad Autónoma de Barcelona. Over the next five years, these databases are likely to double in size, thanks, on the one hand, to the generous, efficient cooperation of national statistical institute partners and, on the other, to major funding by American and European scientific organizations. Although access is restricted to accredited academic researchers and policy-makers, more than 3,000 users representing 76 countries have qualified for access. The IECM initiative has constructed a streamlined harmonization for 35 European census samples, and plans are underway for incorporating more European samples, including microdata for the 2010 round of censuses.

2. The purpose of this paper is to suggest guidelines for preparing census outputs to meet users' needs, specifically census microdata and metadata for timely, efficient processing by academic researchers in general and the IPUMS/IECM projects in particular. We emphasize that over the past decade, the IPUMS/IECM projects have received microdata and documentation in a great profusion of forms. Nonetheless, processing time is reduced and errors minimized when both metadata and microdata are thoroughly documented according to conventional specifications. Statistical offices are strongly encouraged to complete a brief form (see appendix A) describing each set of census microdata and metadata entrusted.

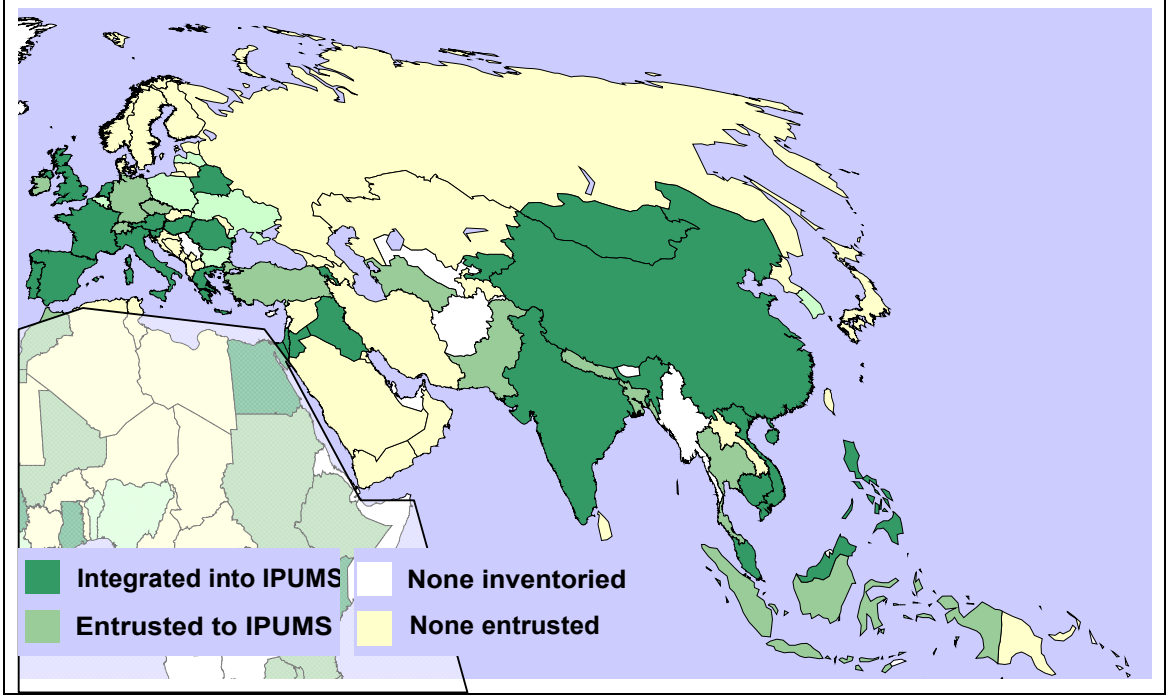
3. Microdata should be transmitted as encrypted executable files, with the password emailed or faxed to the project coordinator in a separate communication. Metadata may be transmitted as images, but should also be made available as ASCII, CSPro, IMPS, NESSTAR, SPSS, STATA, SAS, spreadsheet, or document files, DDI (Data Document Initiative—note that NESSTAR is DDI compliant) hypertext or other emerging standards. Documentation in the official language(s) is essential. English translation should be provided, where available. Otherwise, translators—contracted and paid by the MPC—will prepare unofficial English texts in simple ASCII format. Copies of translations will be provided to the respective National Statistical Office.

II. IPUMS-INTERNATIONAL: "BEST PRACTICE".

4. The epigraph written by Dennis Trewin sums up his glowing assessment of the IPUMS-International facilities, policies and procedures for processing and disseminating anonymized census microdata samples at the Minnesota Population Center (MPC). As the chair of the UNECE task force to produce guidelines on good practice on the release of microdata and the protection of confidentiality, Mr. Trewin is widely recognized as an authority in this field. His strongly positive evaluation of the data protections afforded by the IPUMS-International project provides assurances as we begin our second decade. Readers unfamiliar with the IPUMS-International project's data protections and confidentiality measures are referred to our paper for the UNECE/Eurostat work session on statistical data confidentiality (www.unece.org/stats/documents/2005.11.confidentiality.htm see wp.5) subsequently published in Monographs of Official Statistics (McCaa and Esteve, 2006).

5. 130 anonymized, integrated high-precision samples of population census microdata are presently available at no cost via www.ipums.org/international, the IPUMS-International web-site. Samples for a subset of 12 European countries are available from the Integrated European Census Microdata portal (www.iecm-project.org). These databases are likely to double in size over the next five years, thanks to renewed major funding through 2014 by the National Science Foundation and National Institutes of Health (USA) and to the generous, efficient support of national statistical office partners. More than 3,000 researchers representing 76 countries are accredited to access microdata through these sites. Researchers use these data for comparative analysis across time and space. It is important to note that these projects do not disseminate official statistics. Would-be users seeking suthoritative census figures are directed to websites of the National Statistical Offices, the United Nations Statistical Division and other official sources.

Table 1. IPUMS-EurAsia: Status of Microdata by Country



6. This massive data infrastructure encompasses 44 countries (see Table 1), including for Europe and Asia: Armenia, Austria, Belarus, Cambodia, China, France, Greece, Hungary, Iraq, Israel, Italy, Jordan, Kyrgyz Republic, Mongolia, the Netherlands, Palestine, Philippines, Portugal, Romania, Slovenia, Spain, the United Kingdom, and Vietnam. The IPUMS-International database totals more than 279 million anonymized, integrated person records representing 77 million households. The 2010 release is scheduled to incorporate samples for four EurAsian countries—Nepal, Pakistan, Switzerland, and Thailand—and six from Africa and the Americas—Cuba, Mali, Peru, Saint Lucia, Senegal and Tanzania. Over the next five years we propose to incorporate household samples from the 2010 round censuses as well as microdata from additional countries, such as Bangladesh, Belgium, Bulgaria, Czech Republic, Fiji Islands, Germany, Indonesia, Ireland, Poland, Turkey, Turkmenistan and Ukraine, among others.

7. In the Americas, integration is complete for a dozen countries (56 censuses) with another dozen to be integrated over the next five years. For Africa, barely 13 African censuses (7

countries) are integrated and available for dissemination at this time, but we expect the pace to accelerate soon (McCaa, Esteve, Ruggles and Sobek 2006).

III. THE IECM PROJECT, A EUROPEAN FLAVOURED INTEGRATION

8. The Integrated European Census Microdata project (IECM) was started in 2005 as a joint collaboration between the Centre for Demographic Studies, the Minnesota Population Center and 17 European statistical offices to coordinate, harmonize and disseminate integrated European Census microdata. Closely modeled after the IPUMS-Europe project, the IECM represented the first regional initiative to involve European research and official organizations in the harmonization of census microdata and served as benchmark for similar initiatives in other regions of the world. IECM funding came from the Research Infrastructures action of the European Union Sixth Framework Program. Funding was secured through competitive calls that resulted in three interrelated projects, CIECM, DIECM and HIECM, that ran from 2005 to 2009. (In these acronyms, C stands for coordination, D for dissemination and H for harmonization.) During this period, the IECM team was responsible for coordinating tasks in Europe to marshal the expertise of representatives of national statistical offices and census experts. As a result, four meetings or workshops were held: Barcelona 2005, Paris 2006, Lisbon 2007 and Barcelona 2008.

9. In addition, the IECM team carried out extensive work on the intra-European harmonization of census microdata and developed its own website for dissemination (www.iecm-project.org), from which users have access to the European census microdata (identical to the microdata available on the IPUMS-International website including the intra-Europe variables developed by the IECM project) and to the census documentation provided by the National Statistical Offices, as well as additional Europe-specific metadata not currently available in the IPUMS-International website. Since its inauguration in 2008, the IECM website has provided access to 35 samples from 12 countries from the 1960s to the 2000 rounds of censuses (see Table 2). These samples amount to more than 45 million person records, 14 million households, 166 harmonized variables and 11 intra-European variables.

Table 2. IPUMS-Europe/IECM: Integrated Samples (June, 2009)

Country	Census Year	Sample %	Households (N)	Persons (N)
Austria	1971	10	264,655	749,894
	1981	10	283,693	756,556
	1991	10	310,099	780,512
	2001	10	341,035	803,471
Belarus	1999	10	385,508	990,706
France	1962	5	748,917	2,320,901
	1968	5	815,699	2,487,778
	1975	5	915,624	2,629,456

	1982	5	969,632	2,631,713
	1990	4.2	949,893	2,360,854
	1999	5	1,219,323	2,934,758
Greece	1971	10	249,350	845,483
	1981	10	294,323	923,108
	1991	10	320,387	951,875
	2001	10	367,438	1,028,884
Hungary	1970	5	172,831	515,119
	1980	5	211,355	536,007
	1990	5	219,389	518,240
	2001	5	227,252	510,502
Italy	2001	5	1,168,044	2,990,739
Netherlands	1960	1.2	n.a.	143,251
	1971	1.2	n.a.	159,203
	2001	1.2	n.a.	189,725
Portugal	1981	5	179,409	492,289
	1991	5	214,155	491,755
	2001	5	258,843	517,026
Romania	1977	10	619,904	1,937,021
	1992	10	728,846	2,238,578
	2002	10	732,016	2,137,967
Slovenia	2002	10	63,637	179,632
Spain	1981	5	n.a.	2,084,221
	1991	5	592,276	1,931,458
	2001	5	714,473	2,039,274
United Kingdom	1991	1	215,761	541,894
	2001	3	n.a.	1,843,525
Total			**14,753,767**	**45,193,375**

10. As additional countries participate in this initiative, the IECM team will continue to contribute to the IPUMS projects, subject to new funding opportunities. The EU 7 Framework Program has launched a funding scheme similar to the FP6, in which the IECM team will compete. However, as a research center, and given the consolidation of the current infrastructure,

major effort will be committed to increasing usage and demonstrating the research potential of census microdata. In this regard, several initiatives are underway, such as census microdata training in the European Doctoral School of Demography and research applications based on census microdata (Esteve, et. al. 2009). Dr. Esteve, the IECM research coordinator, has received major funding from the European Research Council to exploit the full capacity of census microdata to investigate the determinants of family life on a global scale. In future years, we expect new projects as analysts incorporate census microdata in their research agenda. The Institut National d'Etudes Démographiques Census Microdata Project, lead by Prof. Patrick Festy, is a noteworthy example. Regarding dissemination, the IECM team continues to exert considerable effort to increase awareness of the data. A concerted campaign was launched for the recent European Population Conference, held in Barcelona in July, 2008. Several sessions were devoted to the project, including a user session with the first research based on European integrated census microdata.

IV. NEED FOR SUCCINCT DESCRIPTIONS OF CENSUS AND MICRODATA: FORM "A".

11. The IPUMS/IECM are bold agendas—indeed, not too long ago, these would have been considered overly ambitious goals, impossible to attain. Even now, if we are to succeed, increased cooperation with our national statistical office partners is essential. As academics, we understand that official statisticians are typically over-burdened with insistent demands from government, business, and the public for an ever increasing array of timely statistics. Therefore we are prepared to work, as we have over the past decade, with metadata and microdata in whatever form without special treatment or consideration. Nonetheless, the integration process is enhanced and errors minimized when both metadata and microdata entrusted are documented according to orderly specifications.

12. Good use requires succinct, authoritative documentation in a readily accessible format. Form "A" (see Appendix A) should be used to succinctly describe the census and microdata. It should be completed by a census expert of the respective National Statistical Office for each set of microdata entrusted. An example of completed forms for the 1981, 1991 and 2001 censuses of Spain is reproduced as Appendix B. Additional examples may be viewed at https://international.ipums.org/international/samples.shtml by clicking the name of a country or at http://www.iecm-project.org/ (first click "Metadata").

13. Form "A" is organized into four categories: description of the census, characteristics of the sample, units identified in the microdata and unit definitions.

(a) Description of the census. The following elements are requested: official title, agency, population universe, de jure or de facto, census day, field work period, number and type of enumeration forms, type(s) of field work, respondent and coverage;

(b) Characteristics of the sample: source (usually the National Statistical Office, National Data Archive or University Research Organization), sample design, sample unit, sample fraction (for both private households and group quarters because these may differ—see below), sample size (number of person records), and a brief description of sample weights;

(c) Units identified in the microdata (indicate yes/no and add any comments desired): dwellings, vacant dwellings, households, individuals, group quarters, lodging, smallest identified geographical unit (name and NUTS), and settled/unsettled/special populations;
(d) Unit definitions: dwellings, private households, group quarters, and settled/unsettled or special populations.

14. Additional items may be added to the form as necessary (e.g., modules regarding emigration, fertility, health insurance, etc.). Where the expert is not fluent in English, the form should be submitted to the MPC in the official language for translation by a professional. If form "A" is already posted on the IPUMS-International website for the country of your expertise, please check entries for each census to confirm that the information is correct and email any suggestions, corrections or comments to ipumsi@pop.umn.edu.

V. METADATA NEEDS

15. Metadata serves a number of purposes within the IPUMS-International/IECM systems. Much of the basic metadata is required to accurately process and assess the microdata as they are incorporated into the databases and to support the harmonization work conducted on specific variables. Comprehensive and complete metadata is essential if the integration is to succeed and researchers are to make informed use of the microdata (Statistics Canada 2008; see also McCaa and Thomas 2009). Metadata may be transmitted as images, but should also be made available as ASCII, CSPro, IMPS, NESSTAR, SPSS, STATA, SAS, spreadsheet, document, or hypertext files or as more than one version of these. When documents are *not* available in electronic form, they are scanned, organized by country and census year, and placed on the IPUMS-International and IECM websites so that they are easily accessible. Copies of census documentation scanned by the MPC or UAB are also made available on CD/DVD to the respective statistical agency as well as national and international research organizations.

16. We have three goals with respect to metadata. First, researchers must have ready access to the original census documentation in the official language. At a minimum, census questionnaires, enumerator instructions or training manuals, and codebooks are required. Additional metadata regarding the organization, preparation work, and actual census taking is also valuable to the IPUMS-International project and is catalogued and archived with all other documents received. Original hardcopy or PDF documents are preferred for published metadata materials. Our goal is to provide an archived collection of high-quality PDF files for all forms of metadata pertaining to census microdata.

17. Census outputs of the following metadata are requested from the National Statistical Offices:

(a) Census enumeration forms;
(b) Census enumerator instructions (sometimes referred to as training manuals);
(c) "Codebooks" or "Data Dictionaries" for each dataset (definitions of record structures, column location of variables and labels for codes, such as the U.S. Census Bureau "IMPS" data dictionary files), including administrative geography, occupations, etc;

(d) Correspondence tables indicating the equivalence between coding schemes in two or more censuses or between a census and an international standard (occupation, education, etc.) These tables are especially helpful to resolve changes in administrative geography and in the integration of educational attainment variables;
(e) Basic tables of official results as they are published on a website, book, or CD.
(f) Technical and methodological reports on census operations, concepts, nomenclatures, comparability, quality, post-enumeration surveys, etc.;
(g) Where microdata are provided in sample form, the sample design should be described in detail. Where the standard IPUMS-International design of every nth household after a random start is employed, no additional documentation is needed (see microdata specifications below). Otherwise, it would be helpful to receive estimates of sampling errors for a scale of absolute or relative frequencies (for example, where sample frequency = 2, 5, 10, 15, 20, 25, 30, 35, 40, and 50 percent), or key variables (or both), such as age, relationship to reference person, education, and employment status. It should be noted that, to date, the National Institute of Statistics of Mozambique has provided the most comprehensive documentation on sample design and errors (Megill 2007);
(h) Boundary files corresponding to the administrative geography coded in the microdata (NUTS1, NUTS2 and NUTS3) and suitable for dissemination to researchers. If boundary files are not provided, we plan to construct unofficial files from readily available sources.

18. Second, we construct a dynamic metadata system for every variable, both integrated and non-harmonized, to make it easy to compare both the phrasing of a particular question and the corresponding instructions to the enumerators, in English, for any combination of countries and censuses. For example, if the researcher wishes to examine the phrasing of the economic activity question, begin by selecting the desired countries and census years. From the home page, click "Variables", then "Select samples." Once the desired samples are selected by ticking a box with the mouse, "submit sample selections," then "mouse-over" "person," scroll down to click "work", then click "EMPSTAT" (the title "Employment Status" is also displayed) , and finally click "Enumeration text." The English text for the selected question and censuses appears on the screen, including both the phrasing on the questionnaire as well as the instructions to the enumerators. Scroll down to easily study every nuance of the source text for each of the censuses. To study a different question for the same set of censuses, backspace or jump to the variable screen and select another variable and click "Enumeration text".

19. Third, from the original source documentation, we write original metadata describing each integrated variable as follows:

(a) brief definition and description of the selected variable,
(b) availability (list of countries and census years with the variable),
(c) general comparability (nuances of varying definitions),
(d) universe (population to which the question is addressed),
(e) reference period (e.g., for economic activity, seven days, last month, a year, etc.),
(f) variations in definitions of specific attributes (e.g., "employed"), and
(g) comparability discussions for specific censuses organized by country.

20. The researcher views these pages by simply clicking the variable name, as described in the previous paragraph. The pages are constructed on demand by the dynamic metadata system. Only the comparability discussions for the currently selected censuses are displayed.

21. Although electronic copies of metadata are preferred, publications or photocopies are also welcome. Electronic files may be emailed as attachments or sent on CDs. Where English translations are needed, professional translators will be contracted and unofficial translations produced in simple text format. To avoid loss of materials and to economize effort, the entire collection should be assembled in a single package, and sent by courier mail at project expense. For participating European countries, examples of original source documentation are posted on the IECM webpage. A nearly complete collection of electronic images of each type of document is posted at the IECM site.

22. For structured metadata (data dictionaries, code lists, definitions, forms, etc.) the use of emerging standards such as the Data Documentation Initiative (www.icpsr.umich.edu/DDI/codebook/) found in NESSTAR and the Microdata Toolkit developed by the International Household Survey Network (http://www.surveynetwork.org/home/) and WorldBank, facilitates the transfer of information into the IPUMS-International processing system. DDI is a mark-up structure using Extensible Markup Language (XML) which identifies specific elements commonly found in the codebook accompanying a data file. It covers identifying information on the data file, census or survey characteristics, sample characteristics, unit definitions, methodology, file structures, variable content and structure, question content and relationship to variables, code lists, and related materials either in-line or through reference to external documents.

23. New versions of DDI, available since 2008, expand coverage to support capturing and relaying information about the complex harmonization process used to construct integrated variables. Soon, we expect to offer to accredited researchers who request microdata extracts the corresponding customized codebooks constructed from the metadatabase underlying the IPUMS-International interface and extraction system. These DDI codebooks could work directly with software such as NESSTAR and the Microdata Toolkit that support DDI documentation as well as provide detailed information not provided within the standard metadata contents of statistical packages. We expect that when DDI compatability is implemented, it will provide a flexible non-proprietary structure for moving metadata into and out of the IPUMS-International system.

VI. MICRODATA NEEDS

24. For microdata we have two goals: first, to permanently archive original source files on behalf of the National Statistical agency partner, and second, to disseminate high-precision, anonymized, integrated and customized household sample extracts to accredited researchers. We prefer that each National Statistical Office entrusts a confidentialized copy (names, addresses, and identification numbers suppressed) of the complete source file (i.e., 100 per cent microdata) so that we may draw the sample consistently, efficiently, and with a minimum of burden on statistical agency partners. Moreover, should there be imperfect records in the sample, any such problems may be resolved easily by replacement, rather than imputation. It should be noted that all microdata source files entrusted to the Minnesota Population Center are archived under total security (“Icebox”) and are never reproduced for any person or institution under any

circumstances. As the Trewin report notes the Minnesota Population Center seeks to maintain a perfect, unblemished record of security.

25. It must be noted that no European statistical office entrusts complete microdata source files. Nonetheless, considering the world as a whole, four modalities for entrusting microdata have emerged over the first decade of IPUMS-International partnerships (bulleted items are examples):

(a) The task of archiving complete source files and producing samples is entrusted to the Minnesota Population Center (38 countries).

(b) Samples produced entirely by the national statistical office according to IPUMS-International specifications where 100% microdata are available (25 countries).
   (i) Federal Statistical Office—Germany: All work performed by FSO, including the 1970 and 1987 censuses of the Federal Republic of Germany and the 1971 and 1981 censuses of the German Democratic Republic.
   (ii) Statistics Netherlands (SN). 1960 and 1971 and a register based sample for 2001—all worked performed by SN.
   (iii) Federal Statistics Office (FSO)—Switzerland: 1971, 1981, 1991, and 2001 – prepared by the FSO.

(c) Public or restricted use microdata samples entrusted to researchers are also entrusted to IPUMS-International with or without payment of license fee (12 countries):
   (i) National Bureau of Statistics, China (license fee paid for 1982; not 1990)
   (ii) National Statistical Survey Organization, India
   (iii) Statistics Canada (no license fee invoiced)
   (iv) Office of National Statistics, United Kingdom (no license fee invoiced)

(d) The task of producing anonymized samples is entrusted to an institution or individual expert under supervision of the national statistical authority (6 countries)
   (i) INSEE—France: 1962, 1968, 1975, 1982, 1990 and 1999 – prepared by an individual researcher working within the INSEE under contract with the Minnesota Population Center and with INSEE oversight.
   (ii) INSSE—Romania: Work performed by a university research institute for the censuses of 2002, 1992, and 1977 under contract with the MPC and with INSSE oversight.

26. Regardless of modality, the project offers a fee of US$5,000 to license microdatasets numbering 1 million or more person records as well as to offset the costs of assembling microdata and documentation. Of course, each national statistical office determines the modality to be used and the project is always amenable to considering other arrangements.

27. A ten percent sample of households is the most common definition of “high precision” (70 of 130 datasets currently integrated), followed by 5 per cent (n=28). Of the 32 samples that are less than 5 per cent, eight are historical samples and include all extant microdata. Where 100% microdata cannot be supplied, we prefer systematic random samples according to the following simple protocol:

(a) Sort the microdata files by major and minor administrative divisions down to the census tract level, dwelling, household, family and person,
(b) After a random start, select every n^{th} private dwelling (every tenth for a 10% sample),
(c) For institutional households—or large private households that could be identifiable solely because of their size—after a random start, draw every n^{th} person using the same density as for private dwellings.

28. Systematic random samples capitalize on 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. 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 (Davern, et. al., 2009).

29. One of the major advantages of using census microdata is its geographical power, which allows sub-national analysis without compromising statistical significance. Due to confidentiality constraints, fine geographical detail must be excluded from census microdata, even when disseminated on a restricted access basis, as in the case of the IPUMS/IECM projects. In spite of the efforts made by the European Union to standardize the administrative divisions of countries for statistical purposes (the NUTS division), EU countries follow different criteria for the geographical disclosure of their census microdata. For comparative purposes, the NUTS3 level would be desirable for all countries (currently only five of 12 EU countries available in IPUMS provide this information) together with additional variables constructed from local administrative units, such as size of place or even individual codes for major metropolitan areas. Providing a size of locality variable would make possible a consistent measure of urban-rural residence across samples. Size of place categories for Germany (preferred) and France are as follows:

	Germany	France
1)	1 to 2,499 persons	1 to 4,999
2)	2,500 to 9,999	5 to 9,999
3)	10,000 to 49,999	10 to 19,999
		20 to 49,999
4)	50,000 to 99,999	50 to 99,999
5)	100,000 to 499,999	100,000 to 1,999,999
6)	500,000 or more	2,000,000 or more

30. Anonymization may be performed by the statistical office or, upon request, by the Minnesota Population Center. Microdata extracts are disseminated to accredited researchers under strict legal and administrative controls (McCaa and Esteve 2006; McCaa, Ruggles, et. al. 2006). While we concur with Anderson and Fienberg (2001) that sampling of datasets alone "provides the additional uncertainty needed to protect many data releases…," we do not stop here. We employ six layers of technical protections. First, we suppress place of enumeration, residence, work or schooling codes for geographical units that fall below a threshold of 20,000

persons in the most recent census. (Some statistical offices set the threshold higher, such as the UK, where the number is 65,000). Second, for categorical variables, any value with a population frequency of less than 250 is likewise suppressed (FSO-Germany is applying a threshold of 2,500). Such values are recoded as either "other," "missing," or in the case of composite codes, the right most digit is coded zero (and the process repeated). Third, for continuous variables, such as income or size of dwelling in square meters, top and bottom coding is used to truncate the tails of distributions as they begin to "thin". Fourth, certain sensitive variables that are particularly susceptible to identifying individuals, such as birth-date, are suppressed. Fifth, a small fraction of households are "swapped" from the geographical unit reported to a neighbouring one to contribute an additional degree of uncertainty. Finally, households are assigned a unique random number and re-sorted.

VII. CONCLUSIONS

31. As census outputs to meet user needs, the IPUMS/IECM projects request a formidable range and amount of metadata and microdata. Nonetheless the return on the investment is substantial Statistical offices are relieved of many of the most burdensome tasks and responsibilities in disseminating microdata to researchers. Moreover, by relying on standard procedures used by a majority of the world's statistical offices, there is safety in numbers. The isolated statistical office that disseminates microdata on an ad hoc basis incurs substantial risks and responsibilities as well as significant human resource and material costs, for a relatively small return with respect to number of users. The IPUMS/IECM projects offer substantial economies of scale with the highest standards of security and disseminates integrated metadata and microdata that greatly facilitates sound scientific research. Participating statistical agencies are invited to entrust metadata and microdata for the 2010 census round at their earliest convenience. Those that are not yet participating in the IPUMS-EurAsia initiative are invited to consider doing so.

References:

Anderson, Margo and Stephen E. Fienberg. (2001). "*U.S. census confidentiality: Perception and reality*," International Statistical Institute Biennial Meeting (Seoul). (unpub.)

CES (2007), "Managing satistical confidentiality and microdata access: Principles and guidelines on good practice", published by the Conference of European of Statisticians: http://www.unece.org/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

Davern, Michael, Steven ruggles, Tami Swenson, J. Trent Alexander and J. Michael Oakes. (2009) "*Drawing statistical inferences from historical census data, 1850-1950*," Demography, 46(3):589-603.

Esteve, A., García, J., Spijker, J., McCaa, R., (2009) "*Integrated european census microdata (IECM) samples: enhancing the study of ageing with high precision over-samples of the oldest-old*", in Work session on statistical data confidentiality, Luxembourg: Eurostat, pp. 407-416.

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

McCaa, Robert; Steven Ruggles, Michael Davern, Tami Swenson, and 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 (New York: Springer), 375-382.

McCaa, Robert and Wendy Thomas. (2009) "*IPUMS-International: lessons from 10 years of archiving and disseminating census microdata*," 57th Session International Statistical Institute, Durban, South Africa (unpub.) www.hist.umn.edu/~rmccaa/ipums-global (scroll to "IPM 100 Microdata Session" and click "IPUMS"

Megill, David. (2007). "*Technical documentation for public use microdata samples files for the 1997 Mozambique census of population and housing*," Instituo Nacional de Estatística, Maputo (unpub.). www.hist.umn.edu/~rmccaa/ipums-africa (click "electronic holdings," scroll to Mozambique, and click "sample design" for the 1997 census.

Statistics Canada. (2008) "*Metadata requirements for archiving structured data*," Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS), Luxembourg.

Trewin, Dennis (2007). *A review of IPUMS-International*. (unpub.).

Appendix A. Form for Recording Census and Sample Characteristics

Instructions: Briefly describe each census and microdata sample.

Text may be in the official language or English. No formatting is required.

Name: _____ email: _____ date: _____

Please check characteristics of other censuses for your country; if integrated, see:

https://international.ipums.org/international/samples.shtml

Address questions to Robert McCaa: rmccaa@umn.edu

Census characteristics (country): _____	
Title	
Census agency	
Population universe	
De jure or de facto	
Enumeration unit	
Census day	
Field work period	
Enumeration forms used	
Type of field work	
Respondent	
Coverage	
Microdata sample characteristics	
Microdata source	
Sample design	
Sample unit	
Sample fraction	
Sample size (person records)	
Sample weights (describe)	
Units identified ("yes" = unit identified; else enter "No")	

Dwellings	
Vacant units	
Households	
Individuals	
Group quarters	
Settled/ Unsettled Population	
Special populations	
Smallest geography in microdata	
Unit definitions	
Dwellings	
Private Households	
Group Quarters	
Unsettled population	
Special populations	
Metadata entrusted (list file names of electronic or titles of paper copies)	
Census forms	
Enumerator instructions/ manuals	
Data Dictionary	
Codebooks (education, occupation, industry, geography, etc.)	
Correspondence tables (education)	
Official results	
Technical, Methodological Reports	
Post-Enumeration Survey Report	
Sample design, sampling errors	
Boundary files (if any)	

Appendix B. Example of Census and Sample characteristics: Spain

https://international.ipums.org/international/sample_designs/sample_designs_es.shtml

Census characteristics			
Census Year	**1981**	**1991**	**2001**
Title	Census of Population and Housing 1981, Spain	Census of Population and Housing 1991, Spain	Census of Population and Housing 2001, Spain
Census agency	Instituto Nacional de Estadisticas (INE)	Instituto Nacional de Estadisticas (INE)	Instituto Nacional de Estadisticas (INE)
Population universe	The census includes persons who have a fixed residence in the national territory as well as those who are in the national territory at the time of reference, without omissions nor duplications. Persons who have their residency in Spain constitute the fixed population [population with rights]. Persons who are in the national territory at the moment of the census form the actual population [de facto].	The census includes persons who have a fixed residence in the national territory as well as those who are in the national territory at the time of reference, without omissions nor duplications. Persons who have their residency in Spain constitute the fixed population [population with rights]. Persons who are in the national territory at the moment of the census form the actual population [de facto].	Residents: persons who at the time of the census habitually reside in Spain. This excludes those persons who were in Spain at the time of the census, but did not reside there. The only exception are citizens and residents temporarily abroad.
De jure or de facto	De facto	De facto	De jure
Census day	1-Mar-81	1-Mar-91	1-Nov-01
Field work period	—	March 1 to April 1, 1991	Two months.
Enumeration forms	Two types of questionnaires are used: one for the population that lives in family dwellings and another for the population that lives in collective dwellings.	There are five forms: general questionnaire, individual questionnaire, collective [dwelling] questionnaire, secondary dwellings with no registered dwellers, and 1990 building questionnaire.	There are four forms: the dwelling, censal data, household data, and individual data forms.

Type of field work	The questionnaires are designed for self-enumeration, so that the task of the Agent is reduced, generally, to the delivering and collecting the questionnaires and verifying the correct completion of the same.	The questionnaires are designed for self-enumeration, so that the task of the Agent is reduced, generally, to the delivering and collecting the questionnaires and verifying the correct completion of the same.	The enumerator is the main form of data collection. However, data can be provided by telephone or internet. The forms are designed to allow the respondents to fill them without any help.
Control of enumeration process	The Group Leaders check the questionnaires one by one; if necessary Agents must go personally to the corresponding dwellings with any incomplete questionnaires, to collect the pertinent information.	—	Enumeration agents collect the information from the different sections (within municipalities) and clarify responses on the forms collected.
Microdata sample characteristics			
Microdata source	Integrated European Census Microdata	Integrated European Census Microdata	Integrated European Census Microdata
Sample design	Systematic stratified sampling. The provinces of Alava, Guipuzcoa, Navarra, and Vizcaya were over-sampled at roughly 5 times the rate of other provinces.	Systematic stratified sampling	Systematic stratified sampling
Sample unit	Dwelling	Dwelling	Dwelling
Sample fraction	5%	5%	5%
Sample size (person records)	2,084,221	1,931,458	2,039,274
Sample weights	Computed by census agency. Use of weights is strongly recommended because of oversampling of several provinces.	Computed by census agency and should be used for most types of analysis.	Expansion factor = 20.
Units identified			

Dwellings	Yes	Yes	Yes
Vacant units	Yes	Yes	Yes
Households	Yes	Yes	Yes
Individuals	Yes	Yes	Yes
Group quarters	Yes	Yes	Yes
Lodging [Alojamientos]	Yes	Yes	Yes
Smallest geography	Municipalities with 20,000+ population	Municipalities with 20,000+ population	Municipalities with 20,000+ population combined by MPC

Unit definitions			
Dwellings	A dwelling is a structurally separate and independent location that, given the way it was built, rebuilt, transformed, or adapted, is conceived to be inhabited by persons, or, even if this is not the case, is effectively and currently inhabited at the time of the census	A dwelling is a structurally separate and independent location that, given the way it was built, rebuilt, transformed, or adapted, is conceived to be inhabited by persons, or, even if this is not the case, is effectively and currently inhabited at the time of the census.	A dwelling is a structurally separate and independent location that, given the way it was built, rebuilt, transformed, or adapted, is conceived to be inhabited by persons, or, even if this is not the case, is effectively and currently inhabited at the time of the census.
Private Households	The family [household] is defined as a group of persons, generally linked because they are relatives, who live together, normally occupying the totality of a dwelling. Persons in domestic services who spend the night in the dwelling and guests who are part of the family group will be included as family.	A household is the group of people who, residing in the same dwelling, share expenses derived from the use of the dwelling and/or meals. Single persons and multiperson households are to be considered.	Group of people residing in the same family dwelling. Sharing of expenses is not required. Family dwellings are those inhabited by one or more persons usually, but not necessarily, having kindred ties and who do not constitute group quarters [see definition below].
Group Quarters	Collective dwellings are those dwellings or buildings designed to be occupied by persons who do not constitute a family, subject to a common regime or authority, or gathered by common objectives or interests.	Collective dwellings are those dwellings or buildings designed to be occupied by persons who do not constitute a family, subject to a common regime or authority, or gathered by common objectives or interests.	Collective dwellings are those dwellings or buildings designed to be occupied by persons who do not constitute a family, subject to a common regime or authority, or gathered by common objectives or interests.
Lodging [Alojamientos]	[Alojamiento] "Lodging" refers to those units that do not qualify as dwelling because they are mobile, semi-permanent or improvised or because they have not been conceived for residential purposes but were occupied during the census.	[Alojamiento] "Lodging" refers to those units that do not qualify as dwelling because they are mobile, semi-permanent or improvised or because they have not been conceived for residential purposes but were occupied during the census.	Lodging is a family dwelling that is mobile, semi-permanent, or improvised or even if it was not intended for residential purposes, is inhabited at the time of the census.