ASSD V: "Information and communication technology in data dissemination: bridging closer producers and users during the 2010 round of Population and Housing Censuses**"**
(19-21 November 2009, Dakar, Senegal)

Timely dissemination of integrated census microdata and metadata: The IPUMS-International approach

Robert McCaa
Minnesota Population Center, Minneapolis, MN USA, rmccaa@umn.edu

Research for this paper was funded in part by the National Science Foundation of the United States, grant SES-0433654 'International Integrated Microdata Series (IPUMS-International year 10)'.

> "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 [on-site] 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), president emeritus International Statistical Institute
> www.hist.umn.edu/~rmccaa/ipums-global/trewin_report_2007.pdf

1. Summary.

1. Integrated, anonymized census microdata and metadata for 130 censuses encompassing 44 countries are presently being disseminated from the IPUMS-International web-site by the Minnesota Population Center (MPC). Africa is represented by only 13 censuses. African statistical institutes are cordially invited to participate in the IPUMS initiative to assure that Africa does not fall behind. Over the next five years, the IPUMS-International database is likely to double in size; however the Africa series is at risk of falling behind because on the one hand too few statistical institutes are participating and on the other there is a rather long delay in entrusting microdata to the project. More than 90 statistical institutes have already endorsed the IPUMS initiative, including all those of America with more than 1 million inhabitants and most of Western Europe.

2. The purpose of this paper is, first, to invite African statistical institutes to participate in the IPUMS-International census microdata project, and, second, to suggest guidelines for bridging the gap between producers and users of census microdata, specifically in preparing census microdata and metadata for timely, efficient processing by academic researchers in general and the IPUMS-International project in particular. Over the past decade, more than 250 sets of census microdata and the corresponding documentation, in a great profusion of forms, have been entrusted to the MPC on behalf of the IPUMS project. Nonetheless, processing time is reduced and errors minimized when both metadata and microdata are thoroughly documented. In addition, statistical institutes are strongly encouraged to complete a detailed form (see appendix A) to accompany each set of census microdata and metadata.

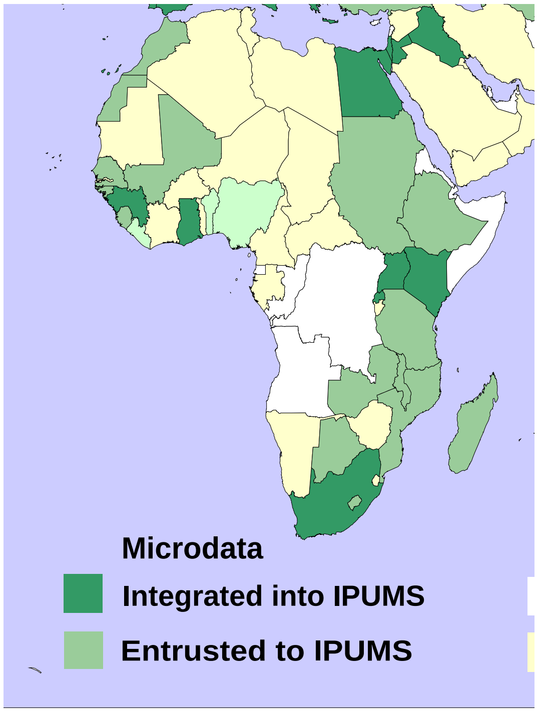
3. For maximum safety, microdata should be transmitted as encrypted executable files, with the password emailed or faxed in a separate communication to the IPUMS-International project coordinator. Metadata may be transmitted as images, but should also be made available as ASCII, CSPro, IMPS, NESSTAR, 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—prepare unofficial English texts in simple ASCII format.

2. Introduction. IPUMS-International: “best practice”.

4. Mr. Dennis Trewin’s accolade “best practice” sums up his meticulous assessment of the IPUMS-International facilities, policies and procedures for archiving, processing and disseminating anonymized census microdata samples. Mr. Trewin, as the chair of the UNECE task force to produce guidelines on good practice for the release of microdata and the protection of confidentiality, is widely recognized as an authority in this field. His strongly positive evaluation of the data protections afforded by the IPUMS-International project assures producers and users alike that we are on the right path as we begin our second decade of activities. 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 subsequently published in Monographs of Official Statistics ((www.unece.org/stats/documents/2005.11.confidentiality.htm see wp.5 and 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. The database is 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 institute partners. More than 3,000 researchers representing 76 countries are accredited to access microdata through the IPUMS-International site. Researchers use integrated census microdata for comparative analysis across time and space. It is important to note that the IPUMS-International project disseminates only integrated, anonymized microdata—not official statistics nor the source files entrusted to the project. Instead, would-be users seeking official census statistics are directed to websites of our National Statistical Institute partners.

Table 1. IPUMS-Africa: Status of Census Microdata by Country (bold year = microdata entrusted)



Country	2000	1990	1980
Samples available to researchers			
Egypt	2006	**1996**	1986
***Ghana**	**2000**		1984
***Guinea, Conakry**		**1996**	**1983**
Kenya	**1999**	**1989**	1979
***Rwanda**	**2002**	**1991**	
South Africa	**2001,7**	**1996**-1	1985-0
***Uganda**	**2002**	**1991**	1980
To be launched in 2010			
***Mali**		**1998**	**1987**
***Senegal**	**2002**		**1988**
***Tanzania**	**2002**		**1988**
In Process			
***Botswana**	**2001**	**1991**	**1981**
***Ethiopia**	2007	**1994**	**1984**
Guinea-Bissau	2009	**1991**	
Lesotho	2006	**1996**	**1986**
Liberia (negotiating)	2008		**1974**
***Madagascar**		**1993**	
***Malawi**	2008	**1997**	**1987**
***Mauritius**	**2000**	**1990**	1983
Morocco	2004	1994	1982
Mozambique	2007	**1997**	1980
Nigeria (negotiating)	2006	1991	
***Sierra Leone**	**2004**		1985
***Sudan**	2008	**1993**	**1983**
***Zambia**	**2000**	**1990**	1980

6. This massive data infrastructure already encompasses 44 countries, including for the continent of Africa: Egypt, Ghana, Guinea (Conakry), Kenya, Rwanda, South Africa and Uganda (see Table 1). 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 three African countries—Mali, Senegal, and Tanzania—plus four EurAsian countries—Nepal, Pakistan, Switzerland, and Thailand—and three American—Cuba, Peru, and Saint Lucia. Over the next five years we propose to incorporate household samples from the 2010 round censuses as well as microdata from other countries, including 13 from Africa (McCaa, Esteve, Ruggles and Sobek 2006). Nonetheless, we must not content ourselves with these twenty-one, otherwise many African countries will remain excluded from enjoying the rewards of participation (countries coloured yellow or white in Table 1).

7. For Africa, barely 13 African censuses (7 countries) are integrated and available for dissemination at this time, accounting for 9.1% of the database (Table 2). This is all the more lamentable because for many countries of Africa, unlike Europe, microdata survive from the 1970s. The IPUMS project is prepared to devote the resources necessary to substantially boost African representation. The first step is for statistical institutes to endorse the standard project memorandum of understanding, such as that signed by the High Commissioner of Planning of Morocco (see Appendix B).

Table 2. IPUMS-Africa: Integrated Samples (September, 2009)

	Census	**Sample**	**Households**	**Persons**
Country	**Year**	**%**	**(N)**	**(N)**
Egypt	1996	10	1,270,787	5,902,243
Ghana	2000	10	397,097	1,894,133
Guinea	1983	10	110,777	457,837
	1996	10	108,793	729,071
Kenya	1989	5	224,861	1,074,098
	1999	5	317,106	1,407,547
Rwanda	1991	10	153,041	742,918
	2002	10	191,719	843,392
South Africa	1996	10	993,801	3,621,164
	2001	10	991,543	3,725,655
	2007	2	345,170	1,047,657
Uganda	1991	10	339,166	1,548,460
	2002	10	529,271	2,497,449

Tallies by Continent

	Samples		
Continent	**Integrated**	**Households**	**Persons**
Africa	13	5,973,132	25,491,624
Americas	56	44,018,013	151,649,996
Asia	26	12,642,375	56,803,289
Europe	35	14,753,767	45,193,375
Total	130	77,469,216	279,464,844

4. Need for succinct descriptions of Census and Microdata: form "A".

8. If the IPUMS-Africa project is to succeed, cooperation of African national statistical institute partners is essential. As academics, we understand that official statisticians are typically over-burdened with pressing 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 by some order.

9. Form "A" (see Appendix A) should be used to succinctly describe each census and its corresponding metadata and microdata. Form A should be completed by a census expert of the respective National Statistical Institute. An example of completed forms for three censuses of South Africa—1996, 2001 and 2007—is reproduced as Appendix C. Additional examples may be viewed at https://international.ipums.org/international/samples.shtml by clicking the name of a country.

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

1) Description of the census. The following elements are requested:
   - i. official title,
   - ii. agency that conducted the census,
   - iii. population universe (note if special populations are omitted, such as nomads, foreigners, etc),
   - iv. de jure or de facto,
   - v. census day(s),
   - vi. field work period,
   - vii. number and type of enumeration forms,
   - viii. type(s) of field work,
   - ix. respondent and
   - x. coverage.

2) Characteristics of the microdata:
   - i. source (usually the National Statistical Institute, National Data Archive or University Research Organization),
   - ii. sample design (preferably every tenth household after a random start),
   - iii. sample unit (household for private entities; individual for collective or group quarters),
   - iv. sample fraction (10% for both private households and group quarters because these may differ—see below),
   - v. sample size (number of person records), and
   - vi. brief description of sample weights, when standard IPUMS protocols are not used.

3) Units identified in the microdata (indicate yes/no and add any comments desired):
   - i. dwellings,
   - ii. vacant dwellings,
   - iii. households,
   - iv. individuals,
   - v. group quarters,
   - vi. lodging,
   - vii. smallest identified geographical unit (name),
   - viii. settled/unsettled/special populations identified in the microdata
   - ix. special household modules (mortality, emigration, agriculture, health, disability, etc.).

4) Unit definitions:
   - i. dwellings,
   - ii. private households,
   - iii. group quarters, and
   - iv. settled/unsettled or special populations.

11. Additional items may be added to the form as necessary (e.g., details for modules regarding mortality, emigration, fertility, agriculture, etc.). The form should be submitted to the MPC in the official language. If form "A" is already posted on the IPUMS-International website for the country of your expertise (see "samples.shtml" web link above), 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.

5. Metadata needs.

12. Metadata serve a number of purposes within the IPUMS-International system. Much of the basic metadata is required to accurately process and assess the microdata as they are incorporated into the database 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 best 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. We are happy to receive more than one version as well. When documents are *not* available in electronic form, the project will scan them, for posting on the IPUMS-International website, organized by country and census year, so that they are easily accessible. Copies of census documentation scanned by the MPC are also made available on CD/DVD to the respective statistical agency as well as national and international research organizations.

13. We have three goals with respect to metadata.

14. 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, data dictionaries and codebooks are required. Additional metadata regarding the organization, preparation, and actual census taking are also valuable to the IPUMS-International project and are 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. Census outputs of the following metadata are requested from the National Statistical Institutes:

1) Census enumeration forms.
2) Census enumerator instructions (sometimes referred to as training manuals).
3) "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.
4) Correspondence tables indicating the equivalence between coding schemes in two or more censuses or between a census and an international standard (ISCO, ISCED, etc.) These tables are especially helpful to harmonize changes in administrative geography and in the integration of occupation, industry, and educational attainment variables.
5) Basic tables of official results as they are published on a website, book, or CD.
6) Technical and methodological reports on census operations, concepts, nomenclatures, comparability, quality, post-enumeration surveys, etc.
7) Where microdata are provided as samples, 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 percent = 2, 5, 10, 15, 20, 25, 30, 35, 40, and 50), and for key variables, 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).

8) Boundary files corresponding to the administrative geography coded in the microdata (corresponding to the European standard of 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.

15. Second, we construct a dynamic metadata system for every variable, integrated as well as 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.

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

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

The researcher views these pages by simply clicking the variable name. The pages are constructed on demand by the dynamic metadata system. Only the comparability discussions for the currently selected censuses are displayed.

17. Electronic copies of source documentation are preferred. Nonetheless, paper publications or photocopies are also welcome. Electronic files may be emailed as attachments or sent by courier service on CDs. Where English translations are needed, professional translators will be contracted and unofficial translations produced in simple text format. To avoid loss of paper or CD materials and to economize effort, the entire collection should be assembled in a single package, and sent by courier at project expense.

18. 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.

19. 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.

6. Microdata needs.

20. For microdata we have two main 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 National Statistical Institutes entrust confidentialized copies (names, addresses, and identification numbers suppressed) of complete source files (i.e., 100% microdata) so that we may draw samples consistently, efficiently, and with a minimum of burden on statistical agency partners. Moreover, should imperfect records be encountered, 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.

21. Additional goals, under consideration, are:

1) Develop an on-line tabulator to offer integrated tabulations for multiple countries and census years. Preferably the tabulator would be harnessed to 100% microdata, but for anonymization purposes, low-level geography would be suppressed. A prototype is already functioning for a dozen European countries.
2) Over-sample important, but infrequently occurring events (maternal mortality) or characteristics (disabilities). For example, from the 100% microdata we propose to include households with all maternal deaths to provide the highest possible precision to analyze this difficult to measure phenomena (see Garenne, McCaa, and Nacro 2008). We have developed a user-friendly method for supplying over-samples without compromising our strong anonymization protections. Moreover our method ensures that researchers use the proper expansion factors.

22. Four modalities for entrusting microdata have emerged over the first decade of IPUMS-International partnerships (bulleted items are examples):

1) The task of archiving 100% microdata source files and producing samples is entrusted to the Minnesota Population Center (38 national statistical institutes).
2) Samples produced entirely by the national statistical institute according to IPUMS-International specifications where 100% microdata are available (25 countries).
   - 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.
   - Statistics Netherlands (SN). 1960 and 1971 and a register based sample for 2001—all work performed by SN.
   - Federal Statistics Office (FSO)—Switzerland: 1971, 1981, 1991, and 2001 – prepared by the FSO.
3) Public or restricted use microdata samples entrusted to researchers are also entrusted to IPUMS-International with or without payment of license fee (12 countries):
   - National Bureau of Statistics, China (license fee paid for 1982; not 1990)
   - National Statistical Survey Organization, India (standard license fee invoiced for 5 samples)
   - Statistics Canada (no license fee invoiced)
   - Office of National Statistics, United Kingdom (no license fee invoiced)
   - Statistics South Africa (no license fee invoiced)
4) The task of producing anonymized samples is entrusted to an institution or individual expert under supervision of the national statistical authority (6 countries)

- 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.
- 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.

23. Each national statistical institute determines the modality to be used. The project is always amenable to considering other arrangements. 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.

24. "High precision" is typically defined as samples of ten percent or higher (70 of 130 datasets currently integrated), followed by 5% (n=28). Of the 32 samples that are less than 5%, thirteen are historical samples and include all extant microdata. Where 100% microdata cannot be entrusted, systematic random samples are preferred according to the following simple protocol:

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

25. 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).

26. 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 project. Typically only the first two levels of geographic detail is provided, such as province and commune, state and county, NUTS1 and NUTS2, etc. In addition, a size of locality variable is preferred because it would facilitate a consistent measure of urban-rural residence across samples. Size of place categories for Germany and France are as follows:

	Germany (preferred)	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 1,999,999	2,000,000 or more
7)	2,000,000 or more	

27. Anonymization may be performed by the statistical institute 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 there. 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 institutes 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 "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, 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 for 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.

7. Conclusions.

28. If we are to bridge the gap between producers and users, new information and communication technologies make census microdata dissemination not only feasible, but easy. The IPUMS project requests a formidable range and amount of metadata and microdata. Nonetheless these are easy to prepare and the return on the investment is substantial By entrusting census microdata to the IPUMS project, statistical institutes are relieved of the far more burdensome, indeed risky, tasks and responsibilities of disseminating microdata to researchers. Moreover, by relying on the standard IPUMS procedures, which are now used by a majority of the world's statistical institutes, 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 project offer substantial economies of scale with the highest standards of security and disseminates integrated metadata and microdata that greatly facilitates sound scientific research. Interactive tabulation of integrated variables offers a vast increase in the number of users and usage of census data with no additional cost to the National Statistical Institute.

29. Statistical institutes participating in the IPUMS-Africa initiative are invited to entrust metadata and microdata for the 2010 census round at their earliest convenience. Institutes that are not yet participating are invited to consider doing so at their earliest convenience.

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 statistical 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.

Garenne, Michel; Robert McCaa and Kourtoum Nacro. (2008) *"Maternal Mortality in South Africa in 2001: From Demographic Census to Epidemiological Investigation,"* Population Health Metrics, 6:4(Aug) 1-13.

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*," Instituto 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.

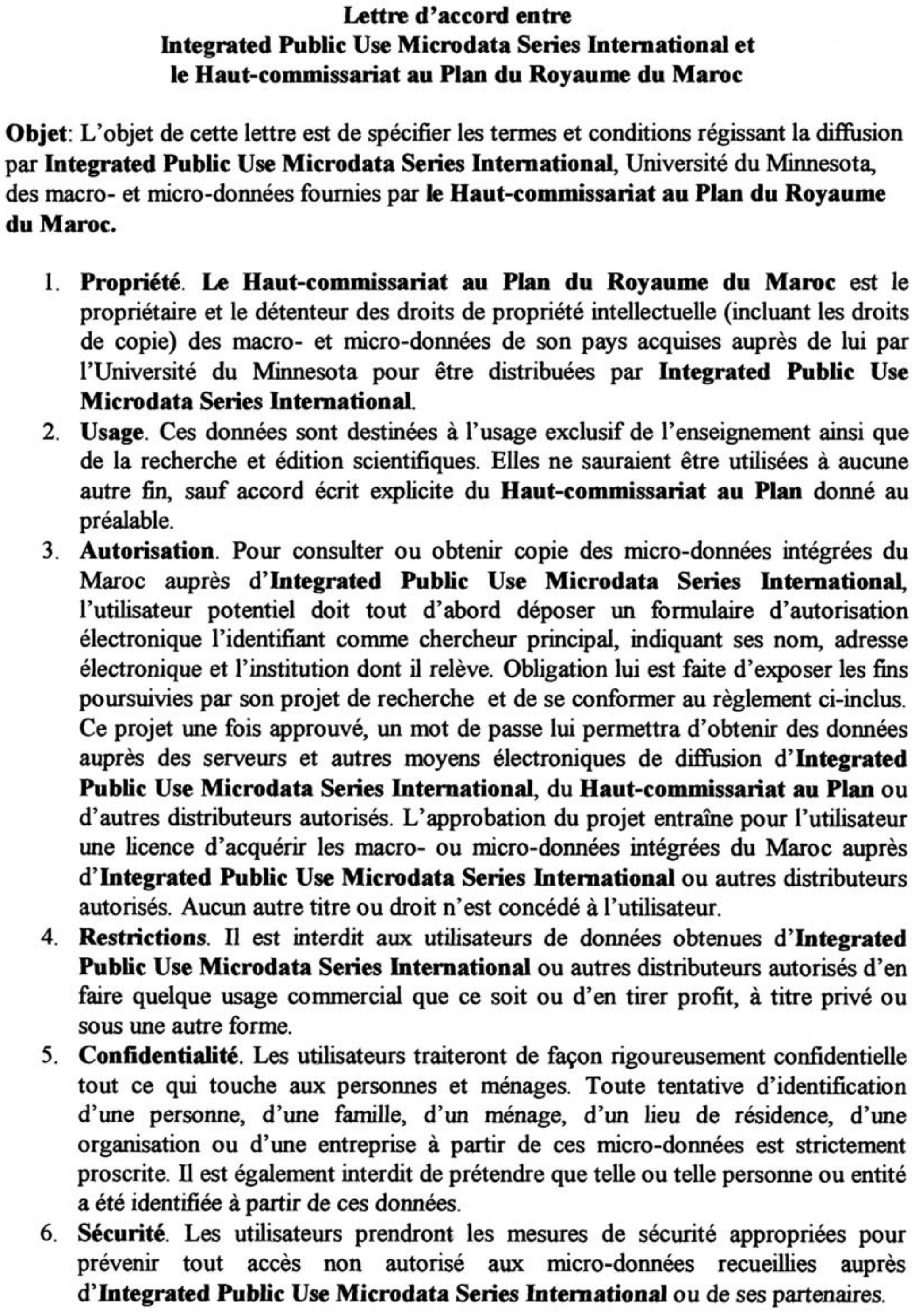
Trewin, Dennis (2007). *A review of IPUMS-International*. (unpub.).

Appendix A. Form "A" for Recording Census and Sample Characteristics

Instructions: Briefly describe each census and microdata sample. 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	
Special modules (mortality, etc.)	
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)	

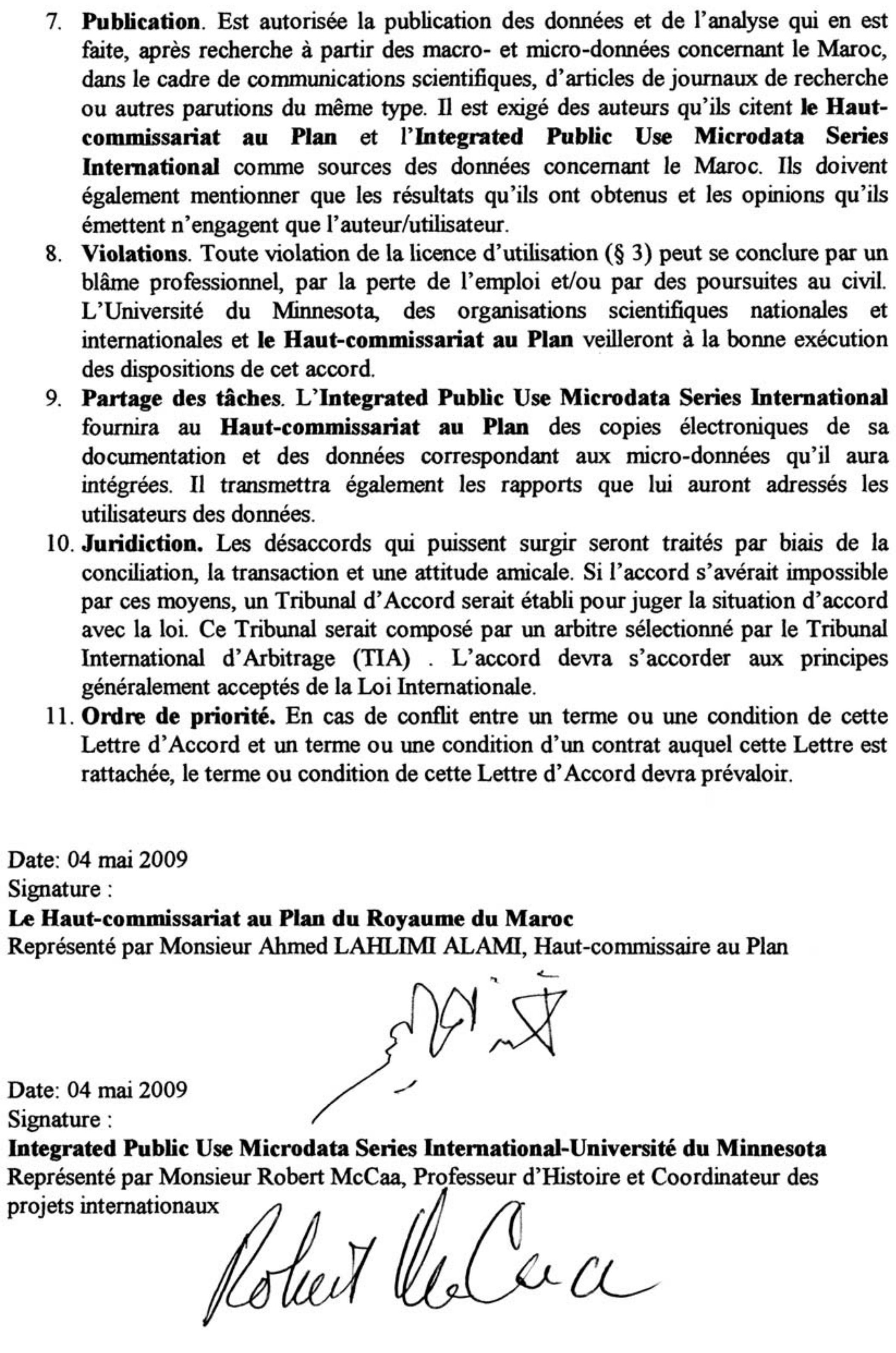
Appendix B. Letter of Understanding: University of Minnesota and the High Commission of Planning of the Kingdom of Morocco



Lettre d'accord entre Integrated Public Use Microdata Series International et le Haut-commissariat au Plan du Royaume du Maroc

Objet: L'objet de cette lettre est de spécifier les termes et conditions régissant la diffusion par **Integrated Public Use Microdata Series International**, Université du Minnesota, des macro- et micro-données fournies par **le Haut-commissariat au Plan du Royaume du Maroc.**

1. **Propriété**. **Le Haut-commissariat au Plan du Royaume du Maroc** est le propriétaire et le détenteur des droits de propriété intellectuelle (incluant les droits de copie) des macro- et micro-données de son pays acquises auprès de lui par l'Université du Minnesota pour être distribuées par **Integrated Public Use Microdata Series International**.
2. **Usage**. Ces données sont destinées à l'usage exclusif de l'enseignement ainsi que de la recherche et édition scientifiques. Elles ne sauraient être utilisées à aucune autre fin, sauf accord écrit explicite du **Haut-commissariat au Plan** donné au préalable.
3. **Autorisation**. Pour consulter ou obtenir copie des micro-données intégrées du Maroc auprès d'**Integrated Public Use Microdata Series International**, l'utilisateur potentiel doit tout d'abord déposer un formulaire d'autorisation électronique l'identifiant comme chercheur principal, indiquant ses nom, adresse électronique et l'institution dont il relève. Obligation lui est faite d'exposer les fins poursuivies par son projet de recherche et de se conformer au règlement ci-inclus. Ce projet une fois approuvé, un mot de passe lui permettra d'obtenir des données auprès des serveurs et autres moyens électroniques de diffusion d'**Integrated Public Use Microdata Series International**, du **Haut-commissariat au Plan** ou d'autres distributeurs autorisés. L'approbation du projet entraîne pour l'utilisateur une licence d'acquérir les macro- ou micro-données intégrées du Maroc auprès d'**Integrated Public Use Microdata Series International** ou autres distributeurs autorisés. Aucun autre titre ou droit n'est concédé à l'utilisateur.
4. **Restrictions**. Il est interdit aux utilisateurs de données obtenues d'**Integrated Public Use Microdata Series International** ou autres distributeurs autorisés d'en faire quelque usage commercial que ce soit ou d'en tirer profit, à titre privé ou sous une autre forme.
5. **Confidentialité**. Les utilisateurs traiteront de façon rigoureusement confidentielle tout ce qui touche aux personnes et ménages. Toute tentative d'identification d'une personne, d'une famille, d'un ménage, d'un lieu de résidence, d'une organisation ou d'une entreprise à partir de ces micro-données est strictement proscrite. Il est également interdit de prétendre que telle ou telle personne ou entité a été identifiée à partir de ces données.
6. **Sécurité**. Les utilisateurs prendront les mesures de sécurité appropriées pour prévenir tout accès non autorisé aux micro-données recueillies auprès d'**Integrated Public Use Microdata Series International** ou de ses partenaires.



7. **Publication**. Est autorisée la publication des données et de l'analyse qui en est faite, après recherche à partir des macro- et micro-données concernant le Maroc, dans le cadre de communications scientifiques, d'articles de journaux de recherche ou autres parutions du même type. Il est exigé des auteurs qu'ils citent **le Haut-commissariat au Plan** et l'**Integrated Public Use Microdata Series International** comme sources des données concernant le Maroc. Ils doivent également mentionner que les résultats qu'ils ont obtenus et les opinions qu'ils émettent n'engagent que l'auteur/utilisateur.
8. **Violations**. Toute violation de la licence d'utilisation (§ 3) peut se conclure par un blâme professionnel, par la perte de l'emploi et/ou par des poursuites au civil. L'Université du Minnesota, des organisations scientifiques nationales et internationales et **le Haut-commissariat au Plan** veilleront à la bonne exécution des dispositions de cet accord.
9. **Partage des tâches**. L'**Integrated Public Use Microdata Series International** fournira au **Haut-commissariat au Plan** des copies électroniques de sa documentation et des données correspondant aux micro-données qu'il aura intégrées. Il transmettra également les rapports que lui auront adressés les utilisateurs des données.
10. **Juridiction.** Les désaccords qui puissent surgir seront traités par biais de la conciliation, la transaction et une attitude amicale. Si l'accord s'avérait impossible par ces moyens, un Tribunal d'Accord serait établi pour juger la situation d'accord avec la loi. Ce Tribunal serait composé par un arbitre sélectionné par le Tribunal International d'Arbitrage (TIA) . L'accord devra s'accorder aux principes généralement acceptés de la Loi Internationale.
11. **Ordre de priorité.** En cas de conflit entre un terme ou une condition de cette Lettre d'Accord et un terme ou une condition d'un contrat auquel cette Lettre est rattachée, le terme ou condition de cette Lettre d'Accord devra prévaloir.

Date: 04 mai 2009
Signature :
Le Haut-commissariat au Plan du Royaume du Maroc
Représenté par Monsieur Ahmed LAHLIMI ALAMI, Haut-commissaire au Plan

Date: 04 mai 2009
Signature :
Integrated Public Use Microdata Series International-Université du Minnesota
Représenté par Monsieur Robert McCaa, Professeur d'Histoire et Coordinateur des projets internationaux

Robert McCaa

Appendix C. Example of IPUMS-International census and sample characteristics metadata: South Africa, 1996, 2001 and 2007

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

	1996	2001	2007
Census characteristics			
Title	Population Census, 1996	Census 2001	Community Survey 2007
Census agency	Statistics South Africa	Statistics South Africa	Statistics South Africa
Population universe	Every person present in South Africa on Census Night, October 9-10, 1996, should have been enumerated.	The people who were present in the country on the night of 9–10 October 2001. People living in households across the country, as well as those in hostels, hotels, hospitals and all other types of communal living quarters, and even the homeless, were all visited.	All usual members of the household who stay in the dwelling at least four nights a week and have done so over the last four weeks prior to census date, plus visitors who spent the night before the interview with the household.
De jure/de facto	De facto	De facto	De facto
Enumeration unit	Visiting points within Enumeration Areas (EAs)	Households and Individuals within Enumeration Areas (EAs- usually contain 100 to 250 households).	Household
Census day	October 10, 1996	October 10, 2001	February 7, 2007
Field work period	October 10 to October 30, 1996 although, in some situations, it was necessary to continue enumeration through to December to ensure that as many people as possible were included.	—	A three week period (February 7-28, 2007) with a non-response follow-up period of one week (March 1-7).
Enumeration forms	There were five different questionnaires that were used: 1) A household questionnaire which was completed in each household in the country. This questionnaire included information on each individual in the household, for example age and gender, as well as on the household as a whole, for example access to electricity and tap water. 2) An individual questionnaire, which was completed by individuals living on their own, for example those living in hostels or compounds. 3) A summary questionnaire for hostels. 4) A questionnaire for institutions, for example prisons, tourist hotels or homes for the aged. 5) A questionnaire for the homeless.	Three different census questionnaires were developed – one for households (the A questionnaire), one for individuals in institutions (the B questionnaire), and one for the institutions themselves (the C questionnaire). The A questionnaire was also used in workers' hostels, student hostels, residential hotels and homes for the independent aged, whilst the B and C questionnaires were also used in tourist hotels and for the homeless.	A single "Household Questionnaire" for information on dwelling, household, and individuals.
Type of field work	Respondents were given the choice of being interviewed or of completing the questionnaire themselves. The vast majority of people chose to be	Each enumerator was required to produce one or more completed questionnaires for each dwelling visited. Households were encouraged to	Direct interview

	interviewed. A general enumerator was a temporary Stats SA staff member appointed to collect information about people who were living in households in private accommodation, for example, a house, a flat in a block of flats, a shack or a traditional dwelling, on census night. A special enumerator was appointed to enumerate people in special dwellings (institutions) such as hostels, prisons, hotels and hospitals. Special enumerators also collected information on the homeless or those living on the streets without shelter or in the open.	be interviewed by the enumerator if possible. Alternatively, a respondent could complete the questionnaire for collection later, where circumstances allowed. Enumerators carried translations of the questions into the other ten official languages, to refer to where necessary.	
Respondent			The head of household or the acting head of the household, and the oldest responsible household member if the head or acting head is not present.
Undercount	10.7%	Varies by province: 14.07% to 22.51% for individuals; 15.55% to 26.21% for households.	Collective living quarters (institutions) and some households in enumeration areas classified as recreational areas or institutions
Coverage			Private dwellings and private seasonal dwellings/holiday homes; workers' hostels and convent/monastery/religious retreats, but not other collective living quarters
Microdata sample characteristics			
Microdata source	Statistics South Africa	Statistics South Africa	Statistics South Africa
Sample design	The household was basically drawn as a 10% systematic sample of households from the census household file, stratified as specified below. The 10% person level sample was obtained by including all persons in these households plus the persons drawn in independent 10% systematic samples of all persons in special institutions and hostels. NOTE: 19 districts in the Eastern Cape province are not organized into households, because of an error in the original data file. 1.3% of the sample is affected.	Systematic stratified sample drawn by Statistics South Africa.	Two-stage, stratified systematic random sampling drawn by the country. Stage I: The sampling frame contains 79,466 enumeration areas (EAs) (the primary sampling units) which are stratified by municipality. Systematic random sampling is used to select EAs within municipalities. In municipalities with fewer than 30 EAs, all EAs are automatically selected. In municipalities with 30+ EAs, a fix proportion of 19% of EAs are selected. If the selected EAs in a municipality are less than 30, the sample in the municipality is increased to 30 EAs. Stage 2: A fixed proportion of 10% of the

			dwellings in a selected EA are selected. If there are less than 10 dwellings in an EA, the selection is increased to 10 dwelling units. All households within the selected dwelling units are covered. No replacement of refusals, vacant dwellings or non-contacts. Response rate 93.9%.
Sample unit	Households and individuals	Households	Enumeration area, dwelling
Sample fraction	10%	10%	2.2%
Sample size (person records)	3,621,164	3,725,655	1,047,657
Sample weights	Computed by census agency and should be used for most types of analysis. The weight variable is the adjustment factor for undercount (for households or persons as appropriate) multiplied by 10 to inflate the 10% sample to the population.	Computed by census agency and should be used for most types of analysis.	Computed by census agency and should be used for most types of analysis.
Units identified			
Dwellings	No	No	No
Vacant units	No	No	No
Households	Yes	Yes	Yes
Individuals	Yes	Yes	Yes
Group quarters	Yes	Yes	Yes
Smallest geography	Magisterial districts with 20,000+ population in 2001	Magisterial districts and municipalities with 20,000+ population in 2001	Municipalities with 20,000+ population
Unit definitions			
Dwellings	An occupied dwelling was a premises (visiting point or physical address) that was inhabited by one or more households on census night. An occupied dwelling may have been a house, room, flat or apartment, shack, hut, tent, caravan, houseboat, shop, school, etc.	Any structure intended or used for human habitation.	A structure or part of a structure or group of strucutres occupied or meant to be occupied by one or more households
Private Households	A household consists of a person, or a group of persons, who occupy a common dwelling (or part of it) for at least four days a week and who provide themselves jointly with food and other essentials for living. In other words, they live together as a unit. People who occupy the same dwelling, but who do not share food or other essentials, were enumerated as	A group of persons who live together, and provide themselves jointly with food and/or other essentials for living, or a single person who lives alone. (The 'four-night-a-week' criterion for household membership does not apply, as this was a de facto census, that is, people were counted where they were staying on census night.)	A household is a group of persons who live together and provide themselves jointly with food or other essentials for living, or a single person who lives alone.

	separate households. For example, people who shared a dwelling, but who bought food and ate separately, were counted as separate households.		
Group Quarters	A special dwelling is one which is not privately occupied by a household. It is usually an institution such as a prison, hotel, hostel, home for the aged, etc. Also hostels: a collective form of accommodation specifically built during the apartheid era for mine, factory, power station, municipal or other employees.	Living quarters where certain facilities are shared by groups of individuals or households. They include hostels, hotels and institutions.	Collective living quarters or communal living quarters are (1) structually separate and independent places of abode intended for habitation by large groups of individuals or several households. Such quarters usually have certain common facilities, such as cooking and ablution facilities, lounges or dormitories which are shared by the occupants. (2) Lving quarters where certain facilities are shared by groups of individulas or households.
Institution			A particular type of collective living quarters, for people with a common characteristic who are living under a common regime. Examples are: hospital/clinic, frail care center, childcare institution/orphanage, etc.