

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS



Managing Statistical Confidentiality & Microdata Access

PRINCIPLES AND
GUIDELINES OF
GOOD PRACTICE



UNITED NATIONS
New York and Geneva, 2007

Note

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations, concerning the legal status of any country, territory, city or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries.

UNITED NATIONS PUBLICATION

Sales No. E.07.II.E.7

ISBN 13: 987-92-1-116959-1

ISSN: 0069-8458

Copyright © United Nations, 2007
All rights reserved

Explanatory Note

These Guidelines have been prepared at the request of the Conference of European Statisticians (CES) by a Task Force chaired by Dennis Trewin (the Australian Statistician).

The guidelines and Core Principles of Confidentiality and Microdata Access were adopted by the CES plenary session on June 2006 and the CES Bureau in October 2006.

The Guidelines will be a dynamic document in that it will be updated from time to time. In particular, it is anticipated that additional Case Studies will be incorporated.

The electronic version of the guidelines is available at the UNECE Statistical Division's website: <http://www.unece.org/stats/publ.htm>

Comments are always welcome. Comments can be sent to confidentiality@unece.org.

Acknowledgements

This work is the result of the efforts of a Task Force set up by the Conference of European Statisticians (CES). The Task Force members were: Ivan Fellegi (Canada), Otto Andersen (Denmark), Teimuraz Beridze (Georgia), Luigi Biggeri (Italy) and Tadeusz Toczynski (Poland). Dennis Trewin (Australia), the Chairman of the Task Force, has made a significant contribution to the work by drafting the text and incorporating the inputs from countries and international organizations throughout the several consultation stages of the document.

Tiina Luige and Gauri Khanna of the Statistical Division of the UNECE were of great assistance to the Task Force.

Svante Öberg (Sweden) and Heinrich Brünger (UNECE) have also provided considerable assistance to the Task Force during the course of their work.

Several countries have provided case studies to support the Guidelines and their efforts are greatly appreciated.

Finally, the Bureau of the CES has provided constructive guidance throughout this project.

CONTENTS

I.	Introduction	1
II.	Why should national statistical offices support the research community?	3
III.	Core principles	6
IV.	Supporting legislation	8
V.	Methods of supporting the research community	9
VI.	Managing tensions between national statistical offices and researchers	16
VII.	Management issues associated with the release of microdata	20
VIII.	Some special issues	22
Annex 1.	Case studies	25
Annex 1.1.	Legislation to support release of microdata - Australia	26
Annex 1.2.	Legislation to support release of microdata - Finland	29
Annex 1.3.	Data cubes - Netherlands	32
Annex 1.4.	Public use microdata - United States	34
Annex 1.5.	Release of anonymised microdata files (licensed files) - Australia	37
Annex 1.6.	Release of licensed microdata files - Netherlands	40
Annex 1.7.	Release of licensed microdata files - Sweden	44
Annex 1.8.	Remote data access facilities - Canada	47
Annex 1.9.	Remote access facility (for microdata access) - Australia	50
Annex 1.10.	Remote access to microdata files - Denmark	52
Annex 1.11.	Research data centre program - Canada	56
Annex 1.12.	Research data centres – United States	59
Annex 1.13.	Data laboratory arrangements - Netherlands	63
Annex 1.14.	Data laboratory microdata access - New Zealand	67
Annex 1.15.	Data laboratory microdata access - Brazil	70
Annex 1.16.	Microdata laboratory analysis - Italy	75
Annex 1.17.	Managing decision making on confidentiality - Slovenia	78
Annex 1.18.	Managing decision making on confidentiality - Australia	83
Annex 1.19.	Microdata access in the OECD programmes for international student assessment (PISA)	85
Annex 1.20.	Policy on international release of microdata - Australia	88
Annex 1.21.	Management of record linkage projects - Canada	91
Annex 1.22.	Data linking when preparing microdata for research - Sweden	95
Annex 1.23.	Access to anonymized census microdata samples via the IPUMS-International and the Integrated European Census Microdata Websites - University of Minnesota Population Center	98
Annex 2.	Standard terminology used in the guidelines	106

ANNEX 1.23. CASE STUDY¹

ACCESS TO ANONYMIZED CENSUS MICRODATA SAMPLES VIA THE IPUMS-INTERNATIONAL AND THE INTEGRATED EUROPEAN CENSUS MICRODATA WEBSITES - UNIVERSITY OF MINNESOTA POPULATION CENTER

1. Broad description

The case study describes the arrangements for providing international and national access to anonymized census microdata samples via the IPUMS - International and the Integrated European Census Microdata websites (University of Minnesota Population Center and the Centre d'Estudis Demogràfics, Autonomous University of Barcelona) with France, as a specific example.

High precision, anonymized, integrated census microdata are available to researchers on a restricted access basis from IPUMS-International (www.ipums.org/international). Terms are specified by a memorandum of understanding negotiated between each National Statistical Office and the University of Minnesota. This method of dissemination is governed, on the one hand, by legislation requiring that the data be held in strict confidence and used exclusively for statistical purposes and, on the other, by a stringent license agreement between the University of Minnesota and each user. In May 2002, anonymized, integrated microdata samples for the French censuses of 1962, 1968, 1975, 1982 and 1990 were released, along with samples for China, Colombia, Kenya, Mexico, the USA and Vietnam. The December 2006 release includes samples for the censuses of Belarus, Greece, Romania and Spain as well as the Philippines and Uganda. As of January 1, 2007, the database comprises 63 samples, 20 countries, and 185 million person records. An additional six European statistical agencies (and 38 non-European) have provided census microdata to the project: Austria (4 censuses), Czech Republic (2), Hungary (4), Netherlands (3), Portugal (3), and the United Kingdom (2; the 1981 and earlier censuses are under consideration). Five other European countries have endorsed the project, but have not yet provided data: Bulgaria, Germany, Italy, Slovenia, and Turkey. Beginning in 2008, the European microdata will also be distributed by the Integrated European Census Microdata (IECM) project using identical protocols, although the microdata will be harmonized according to European, rather than global, practices.

2. Why is it good practice?

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

¹ This case study is available only in the online version of the publication.

3. Target audience

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

4. Detailed description

The IPUMS-International project is governed by a uniform Memorandum of Understanding (MOU) signed with each participating National Statistical Office (see Box 1). The MOU confirms that the National Statistical Office specifies the terms and conditions under which the microdata and metadata entrusted to the University of Minnesota and the Autonomous University of Barcelona shall be governed:

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

In addition there are statements regarding:

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

The principal sanction for misuse is recall of data and an embargo against use by the individual and the individual's institution. In addition, the sanctions clause of the MOU threatens additional sanctions to assure compliance:

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

4.1 Data confidentiality

Before providing census microdata to the Minnesota Population Center, the National Statistical Office imposes a number of undisclosed technical confidentiality measures. The Minnesota Population Center imposes an additional suite of techniques such that any allegation that an individual has been identified with absolute certainty is false. In addition, to further ensure the confidentiality of the microdata, administrative geography is limited. In the case of France 22 regions are identified. The smallest has a population exceeding 80 000 in the 1990 census (sample $n > 4\ 000$). The sample count for any identifiable single year of

age is >100. For any identifiable country of citizenship the sample count is >100. Each National Statistical Office determines the minimum population threshold for the identification of administrative geography and other sensitive characteristics, such as ethnicity, country of birth, citizenship, etc.

4.2 Rules and procedures regarding release to users

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

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

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

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

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

And:

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

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

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

5. Supporting legislation (example of France)

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

6. Strengths

- a. Offers security against loss of source microdata. Raw data files entrusted to the project are encrypted and stored in a secure data repository. Copies of these files are made available only to the National Statistical Office-owner, and are never re-distributed to others.
- b. Fosters maximum uniformity of approach and facilitates greater access to microdata by the research community.
- c. Improves on arrangements for providing access to microdata to the greater satisfaction of both the National Statistical Offices and the research community.
- d. National Statistical Offices cede census microdata files to the University of Minnesota. The data are anonymized and then integrated. Much new integrated metadata are written and stored in a database accessible to all at no cost via the internet. Integrated microdata are available for dissemination on a licensed basis to approved researchers. All licensed microdata disseminated by the University of Minnesota Population Center are governed by a uniform Memorandum of Understanding (MOU) between the National Statistical Office and the University. If requested to do so, the University will cease dissemination and return all copies of census microdata in its possession to the corresponding National Statistical Office.
- e. Employees of the University who work with original source data are certified in human subject protections, including the protection of statistical confidentiality. Violations are punishable by termination of employment, and, at the discretion of the University, civil prosecution with a maximum fine of US\$ 250 000 and/or three years imprisonment.
- f. The means of gaining access to the microdata are transparent and equitable. They are based on the principle of freedom of scientific inquiry, regardless of country of birth, residence, workplace or citizenship. Decisions to grant access are determined by project principal investigators. Each individual who wishes to work with the microdata is required to be licensed. The license is valid for one year and is renewable. A condition for renewal is the sharing of research findings, which, in turn, are made available to the national statistical offices.
- g. Microdata are available as extracts on a licensed basis only to researchers who agree to abide by the conditions of use and demonstrate a bona fide research need to access the data. The license constitutes a legally binding undertaking. An attempt to match individuals constitutes a violation of the license agreement and would lead to recall of data and sanctions against both the individual and his/her institution.

- h. Sanctions for breaches of the license agreement are clearly spelled out. These include:
 - i. sanctions against both the individual and the institution with which the individual is associated (e.g., University, international organization);
 - ii. denial of access would immediately be invoked against the individual and his/her institution and would continue until corrective measures were deemed to be sufficient by the University of Minnesota and the National Statistical Office whose data were violated. If the institution where the breach occurred was the recipient of a grant from the National Institutes of Health of the United States, each researcher at the institution could be required to undergo Human Subjects Protection training and re-certification before access was re-instituted for individuals at that institution.
 - iii. civil prosecution could be instituted with assistance requested, under the terms of the project MOU, of the National Statistical Office of the country in which the violation occurred to the extent permitted by national legislation.
- i. Microdata are encrypted during transmission using 128-bit SSL (Secure Sockets Layer) encryption standards used by the financial industry.
- j. Anonymization protocols (top coding, bottom coding, grouping of small cell counts, collapsing of variables, randomization of records and some recodes, suppression of sensitive variables, etc.) are rigorous, yet precision of samples is high. Anonymization protocols are determined by each National Statistical Office before extracts of the data are disseminated.
- k. Integrated metadata are provided describing census operations, sample methodologies, variables and codes. The documentation is harmonized so that researchers who become familiar with the metadata for one census will readily understand the metadata system for any other census of any other country.
- l. Microdata consist of high precision household samples with many integrated, value-added variables—such as “WTPER”, which specifies the person weight for each record in every sample; “SUBSAMP”, which provides 100 certified sub-samples which researchers may use to generate robust estimates of sample variance; “SPLOC” which points to the spouse of each individual whose spouse is co-resident in a household; etc.
- m. Costs are borne through sustained funding from the National Science Foundation of the United States of America with supplementary funds provided by the National Institutes of Health. Where required, the project pays a license fee to the National Statistical Office for the documentation and microdata. The fee is intended to cover marginal costs for the National Statistical Office to provide technical assistance in developing the microdata samples and interpreting the documentation. The *European Union Sixth Framework Programme* provides support to the IECM project for enhancing, harmonizing and disseminating the integrated European microdata and metadata as well as for coordinating tasks based in Europe.

7. Weaknesses

- a. National Statistical Offices cede authority to the University to grant access to census microdata extracts to bona fide researchers. Decisions to grant access are determined by project principal investigators.
- b. Microdata are not wholly anonymized. With sufficient resources, in terms of computing power, time, and a companion microdata set, data matching could be performed to identify individuals to a high probability, although not with absolute certainty.

- c. Misuse of microdata by even one researcher may impact negatively on the ability of a National Statistical Office to obtain cooperation of respondents in that country, or even conceivably, other countries.
- d. Users do not have access to original source files supplied by the National Statistical Office. Instead researchers access integrated microdata with codes and documentation which not only may differ from the original source but also may contain errors introduced in the integration process.
- e. Quality of microdata may not be sufficiently high for the intended research purpose.
- f. Whether the license constitutes a legally binding undertaking has not been tested in a court of law.
- g. There is no requirement that the microdata be destroyed once the initial research is completed.
- h. There is no opportunity for the National Statistical Office to comment upon the research before it is published.

8. References

Bruengger, Heinrich. 2004. "The relationship between the fundamental principle on confidentiality and population censuses: Statement from the UNECE Statistical Division," United Nations Symposium on Population and Housing Censuses: New York, September 13-14.

Isnard, Michel. 2006. "Statistics and individual liberties: recent changes in French law," Courrier des statistiques, English series no.12, pp. 26-30.

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

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

McCaa, Robert and Albert Esteve. 2006. "IPUMS-Europe: Confidentiality measures for licensing and disseminating restricted access census microdata extracts to academic users," Monographs of official statistics: Work session on statistical data confidentiality. Luxembourg: Office for Official Publications of the European Communities, pp. 37-46.

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

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

Letter of Understanding

Box 1

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

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

1. **Ownership.** The [National Statistics Institute of Country X] is the owner and licensee of the intellectual property rights (including copyright) in the metadata and microdata of [Country X] acquired by the University of Minnesota to be distributed by **Integrated Public Use Microdata Series International**. This agreement explicitly authorizes release to the University of microdata of [Country X] that may be in the possession of third parties. The University is obligated to provide to the [National Statistics Institute of Country X] timely notice of any such acquisitions and, upon request and without cost, provide copies of same.
2. **Use.** These data are for the exclusive purposes of teaching, scientific research and publishing, and may not be used for any other purposes without the explicit written approval, in advance, of the [National Statistics Institute of Country X].
3. **Authorization.** To access or obtain copies of integrated microdata of [Country X] from **Integrated Public Use Microdata Series International**, a prospective user must first submit an electronic authorization form identifying the user (i.e., principal investigator) by name, electronic address, and institution. The principal investigator must state the purpose of the proposed project and agree to abide by the regulations contained herein. Once a project is approved, a password will be issued and data may be acquired from servers or other electronic dissemination media maintained by **Integrated Public Use Microdata Series International**, the [National Statistics Institute of Country X], or other authorized distributors. Once approved, the user is licensed to acquire integrated metadata and microdata of [Country X] from **Integrated Public Use Microdata Series International** or other authorized distributors. No titles or other rights are conveyed to the user.
4. **Restriction.** Users are prohibited from using data acquired from the **Integrated Public Use Microdata Series International** or other authorized distributors in the pursuit of any commercial or income-generating venture either privately, or otherwise.
5. **Confidentiality.** Users will maintain the absolute confidentiality of persons and households. Any attempt to ascertain the identity of a person, family, household, dwelling, organization, business or other entity from the microdata is strictly prohibited. Alleging that a person or any other entity has been identified in these data is also prohibited.
6. **Security.** Users will implement security measures to prevent unauthorized access to microdata acquired from **Integrated Public Use Microdata Series International** or its partners.
7. **Publication.** The publishing of data and analysis resulting from research using metadata or microdata of [Country X] is permitted in communications such as scholarly papers, journals and the like. The authors of these communications are required to cite [National Statistics Institute of Country X] and **Integrated Public Use Microdata Series International** as the sources of the data of [Country X], and to indicate that the results and views expressed are those of the author/user.
8. **Violations.** Violation of the user license may lead to professional censure, loss of employment, and/or civil prosecution. The University of Minnesota, national and international scientific organizations, and the [National Statistics Institute of Country X] will assist in the enforcement of provisions of this accord.
9. **Sharing.** **Integrated Public Use Microdata Series International** will provide electronic copies to the [National Statistics Institute of Country X] of documentation and data related to its integrated microdata as well as timely reports of authorized users.
10. **Jurisdiction.** Disagreements which may arise shall be settled by means of conciliation, transaction and friendly composition. Should a settlement by these means prove impossible, a Tribunal of Settlement shall be convened which will rule upon the matter under law. This Tribunal shall be composed of an arbitrator, who shall be selected by the ICC International Court of Arbitration. This agreement shall be governed by, and construed in accordance with, generally accepted principles of International Law.
11. **Order of Precedence.** In the event of a conflict between a term or condition of this Letter of Understanding and a term or condition of any Contract, to which this Letter of Understanding is attached, the term or condition in this Letter of Understanding shall prevail.