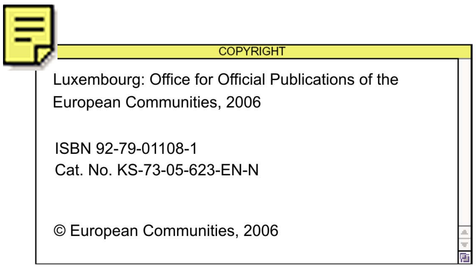
RESEARCH IN OFFICIAL STATISTICS

2006 EDITION

Monographs of official statistics

Work session on statistical data confidentiality

Geneva, 9-11 November 2005



COPYRIGHT

Luxembourg: Office for Official Publications of the European Communities, 2006

ISBN 92-79-01108-1
Cat. No. KS-73-05-623-EN-N

EUROPEAN COMMISSION

UNITED NATIONS
ECONOMIC
COMMISSION
FOR EUROPE

eurostat

THEME
General
and regional
statistics

A great deal of additional information on the European Union is available on the Internet.
It can be accessed through the Europa server (http://europa.eu.int).

Luxembourg: Office for Official Publications of the European Communities, 2006

ISBN 92-79-01108-1

EUROSTAT
L-2920 Luxembourg — Tel. (352) 43 01-1 — www.europa.eu.int/comm/eurostat/

Eurostat is the Statistical Office of the European Communities. Its task is to gather and analyse figures from the different European statistical offices in order to provide comparable and harmonised data for the European Union to use in the definition, implementation and analysis of Community policies. Its statistical products and services are also of great value to Europe's business community, professional organisations, academics, librarians, NGOs, the media and citizens.

To ensure that the vast quantity of accessible data is made widely available and to help each user make proper use of the information, Eurostat has set up a publications and services programme.

This programme makes a clear distinction between general and specialist users and particular collections have been developed for these different groups. The collections *Press releases*, *Statistics in focus*, *Panorama of the European Union*, *Pocketbooks* and *Catalogues* are aimed at general users. They give immediate key information through analyses, tables, graphs and maps.

The collections *Detailed tables* and *Methods and nomenclatures* suit the needs of the specialist who is prepared to spend more time analysing and using very detailed information and tables.

As part of the new dissemination policy, Eurostat has developed its website. All Eurostat publications are downloadable free of charge in PDF format from the website. Furthermore, Eurostat's databases are freely available there, as are tables with the most frequently used and demanded short- and long-term indicators.

Eurostat has set up with the members of the 'European statistical system' a network of support centres which will exist in nearly all Member States as well as in some EFTA countries. Their mission is to provide help and guidance to Internet users of European statistical data. Contact details for this support network can be found on our Internet site.

Eurostat

Table of contents

Acknowledgements

The work session on statistical confidentiality was jointly organized by Ms. Maria João Santos from Eurostat Unit for Methodology and Research and Mr. Juraj Riecan from UN Economic Commission for Europe.

UNECE and Eurostat gratefully acknowledge the valuable contributions of all the participants and particularly the session organizers and discussants, who were the following: Mr. Heinrich Brüngger (UNECE), Mr. Pedro Díaz Muñoz (Eurostat), Mr. Lawrence H. Cox (United States of America), Mr. Josep Domingo-Ferrer (Rovira i Virgili University, Spain), Ms. Luisa Franconi (Italy), Ms. Sarah Giessing (Germany), Mr. Anco Hundepool (Netherlands), Mr. Jean-Marc Museux (Eurostat) and Mr. Eric Schulte Nordholt (Netherlands).
Special thanks to Mr. Anco Hundepool for chairing the session.

The proceedings have been produced by Ms. Maria João Santos and Ms. Aleksandra Bujnowska from Eurostat Unit for Methodology and Research.

Foreword

Statistical confidentiality primarily aims at safeguarding privacy in the field of statistics and is a key to the necessary trust that has to be maintained between statistical bodies and respondents. Mutual confidence ensures accurate and reliable basic information and eventually high quality statistics.

There is a growing appreciation of the benefits of providing access to microdata for research and analysis. At the same time it is vital to protect data confidentiality. It is essential that new approaches are developed at international level to meet these objectives which create conflicting pressures. The risks to confidentiality must be managed effectively. A key challenge is how to minimise the risks to confidentiality, including the perception of threats to confidentiality. Striking the right balance is vital.

The work session covered a wide range of different aspects of statistical confidentiality from remote access to risk management by adequate access procedures to microdata.

The agenda of the work session consisted of the following topics:

(i) Web/on-line remote access (techniques, confidentiality protection and organizational issues);
(ii) Disclosure risk, information loss and usability of data;
(iii) Confidentiality aspects of statistical information taking into account register-based data;
(iv) Access to business microdata for analysis;
(v) Confidentiality aspects of tabular data, frequency tables, etc.;
(vi) Software for statistical disclosure control;
(vii) General statistical confidentiality issues (legal framework, political and conceptual aspects, terminology).

Papers presented under topic (i) focused on 2 types of access: remote execution, which is less flexible but provides better disclosure control and where all outputs are checked; and remote access, which is more flexible but disclosure control is more difficult and final output is checked.

The discussion on topic (ii) focused on the release of microdata files that may lead to risk of disclosure. The participants discussed several methods for assessing disclosure risk, as a crucial element of disclosure control and stressed the importance of statistical models.

In topic (iii) aspects of statistical disclosure control were discussed in the presence of accessible registers and archives that may permit re-identification of records.

Several methods were discussed in topic (iv) for secure computation that may allow sharing business data without compromising data confidentiality. These methods included secure summation protocols, secure matrix product protocols, and synthetic data approaches.

Papers presented under topic (v) discussed several methods to protect tabular data from rounding, peturbative methods such as controlled tabular adjustment or the use of fixed intervals as an alternative to cell suppression.

In session (vi) software solutions covering the entire field of statistical disclosure control were presented, amongst others: method producing safe output for complex statistical analysis in a remote access environment, algorithm for controlled tabular adjustment, SUDA program for classifying cell according to their disclosure risk, use of τ-Argus software for cell suppression.

Papers presented in topic (vii) discussed matters such as the balance that needs to be found between the need to provide users with access to microdata and the need to protect the confidentiality of respondents, legal and administrative procedure as part of risk management, harmonization of SDC methods and procedures on international level and production of data confidentiality and microdata access guidelines.

The work session was a great opportunity for official statisticians and researchers to exchange ideas and discuss new methods and tools dealing with confidentiality. The papers presented hereafter constitute a very important contribution to the development of applied procedures in this domain.

Pedro Díaz Muñoz

Heinrich Brüngger

IPUMS-Europe: Confidentiality measures for licensing and disseminating restricted-access census microdata extracts to academic users

Robert McCaa and Albert Esteve***

*** Minnesota Population Center, Minneapolis, MN 55455, USA,**
rmccaa@umn.edu
**** Centre d'Estudis Demogràfics, Autonomous University of Barcelona, Barcelona, Spain,**
aepalos@yahoo.es

Research for this paper was funded in part by the National Institutes of Health of the United States, grant HD047283'European census microdata harmonization project (IPUMS-Europe)'and CIECM.

"Inadequate use of microdata has high costs"
—Len Cook (2003)

Abstract. Confidentiality protections for census microdata depend not only on the sensitivity and heterogeneity of the data, but also on the potential users. It is widely recognized that statistical agencies exert substantial effort to protect microdata from misuse by academics, their most trust-worthy users. The IPUMS-International projects, by disseminating only integrated, anonymized microdata and restricting access to licensed academic users, shifts the risk-utility curve sharply rightward—substantial increasing utility with only marginal increments in risk. The IPUMS-International approach provides access to microdata of high utility at the same time that confidentiality risks are minimized. Many statistical institute partners anonymize the microdata and implement technical measures of confidentiality protection before the data are entrusted to the project. This paper discusses legal, administrative and technical practices of the IPUMS-International project for disseminating harmonized census microdata extracts with specific reference to the IPUMS-Europe regional initiative.

1. Introduction: IPUMS-International

The IPUMS-International is a global initiative led by the University of Minnesota Population Center to confidentialize, harmonize and disseminate high-density census microdata samples on a restricted access basis to academic users (Ruggles et. al. 2003). Begun in 1999 with funding provided by the National Institutes of Health and the National Science Foundation of the United States, to date the initiative enjoys the endorsement of official statistical institutes of more than fifty countries. Marginal costs of constructing and maintaining the database are born by the MPC, its funding agencies, the University of Minnesota and academic partners–not by the statistical institute partners. On the contrary each is paid a modest fee per census to supply microdata and documentation to the project. In May 2002, the first phase of integrated census microdata for Colombia (1964-1993), France (1962-1990), Kenya (1989-1999), Mexico (1960-2000), the United States (1960-2000), and Vietnam (1989-1999) were made available to licensed users, followed by China (1982) in 2003 and Brazil (1960, 1970, 1980, 1991, 2000) in 2004. More than 500 users representing more than 30 countries are currently licensed to obtain custom-tailored extracts free of charge from the project website: https://www.ipums.org/international

Table 1. IPUMS-International Integrated Census Microdata Sample Characteristics, 120 million person records.

Country census		**Sample %**	**No. of Person records**	**Additional details**
Brazil	1960	5.0	3,001,000	Long-form, cluster sample
	1970	5.0	4,954,000	Same
	1980	5.0	5,870,000	Same
	1990	5.0	8,523,000	Same
	2000	6.0	10,136,000	Same
China	1982	0.1	1,003,000	Every thousandth household
Colombia	1964	2.0	350,000	Every fiftieth person
	1972	10.0	1,989,000	Every tenth household
	1985	10.0	2,643,000	Long-form, cluster sample
	1993	10.0	3,247,000	Every tenth household
France	1962	5.0	2,321,000	Every twentieth household
	1968	5.0	2,488,000	Same
	1975	5.0	2,629,000	Same
	1982	5.0	2,714,000	Same
	1990	4.2	2,361,000	Every twenty-fourth household
Kenya	1989	5.0	1,074,000	Every twentieth household
	1999	5.0	1,410,000	Same
Mexico	1960	1.5	503,000	Every 67th individual
	1970	1.0	483,000	Every hundredth household
	1990	10.0	8,028,000	Every tenth household
	2000	10.6	10,099,000	Long-form, cluster sample
USA	1960	1.0	1,800,000	Stratified, random sample
	1970	1.0	2,030,000	Same
	1980	5.0	11,337,000	Same
	1990	5.0	12,500,000	Stratified, cluster sample
	2000	5.0	14,082,000	Same
Vietnam	1989	5.0	2,627,000	Long-form, cluster sample
	1999	3.0	2,368,000	Same

Source: https://www.ipums.org/international/sample_descriptions.html

With the inclusion of the data for Brazil, the IPUMS-International website offers some 120 million person records consisting of more than 100 variables from 28 samples with densities varying from 0.1 to 10 percent (Table 1). Over the next five years, the database will expand to 44 countries with regional initiatives in Europe, Africa, Asia, Oceania and Latin America (McCaa and Esteve 2005). It should be noted that the mode of access to IPUMS-USA samples differs from the International project. The former, a public site, makes data available to anyone and therefore has tens of thousands of users, while the later provides data only to licensees, numbering only in the hundreds.

The IPUMS-International/Europe regional project began in September 2004. Thanks to additional funding by the European Community Sixth Framework Program, the inaugural workshop was held in Barcelona in July 2005. Delegates from the official statistical agencies and academics met to discuss data availability, samples, general harmonization issues, and overall project procedures. A second workshop, hosted by L'Institut National d'Études Démographiques, to be held in June, 2006 will focus on detailed harmonization issues. In 2007, the first European region data release is scheduled for release with a mirror-site at the Centre d'Estudis Demogràfics (Barcelona).

Table 2. IPUMS-Europe: Likely Censuses and Sample Sizes (in 000s), by Country. Bolded census year indicates sample has been drawn and entrusted to project.

	Sample Density (%)	Census	N	Census	N	Census	N	Census	N	Census	N
Austria	**10**	**2001**	**884**	**1991**	**902**	**1981**	**838**	**1971**	**836**		
Belarus	**10**	**1999**	**1,040**			**.**	**.**	**.**	**.**	**.**	.
Bulgaria	5	2001	395	1992	425	.	.	.	.	.	.
Czech Rep.	5	2001	515	1991	515	.	.	.	.	.	.
France	**5**	**1999**	**3,005**	**1990**	**2,361**	**1982**	**2,714**	**1975**	**2,629**	**1968**	**2,488**
Germany	?	2001	330	1991	321	1987	tbd	1982	tbd	1973	246
Greece	**10**	**2001**	**1,029**	**1991**	**969**	**1981**	**923**	**1971**	**845**	**.**	**.**
Hungary	**5**	**2001**	**511**	**1990**	**518**	**1980**	**536**	**1970**	**515**	**.**	**.**
Netherlands	**1**	**2001**	**190**	**.**	**.**	**.**	**.**	**1971**	**159**	**1960**	**143**
Poland	?	2002	1,930	1995	1,940	1988	1,900	1984	1,850	1978	1,745
Portugal	?	2001	500	1991	495	1981	490	.	.	.	.
Romania	**10**	**2002**	**2,239**	**1992**	**2,138**	.	.	.	.	.	.
Russia	5	2002	7,200	1989	7,400	.	.	.	.	.	.
Slovenia	10	2001	200	1991	tbd			.	.	.	.
Spain	**5**	**2001**	**2,040**	**1991**	**1,940**	**1981**	**1,875**	.	.	.	.
UK	1	2001	600	**1991**	**574**	.	.	.	.	.	.

Notes: Total Person Records ~ 65 million

Micro-censuses: Germany 1982, 1991, 2000; Netherlands 2001; Poland 1974, 1984, 1995.
Samples for 1962 France and 1960 and 1974 Poland are included in the total case count.
Final agreements for Poland, Russia and Turkey are pending, and some of the earliest censuses may not be recoverable.
tbd = to be determined.

2. Dissemination of IPUMS-International "Extracts"

Users of IPUMS-International are not permitted to access microdata containing the original codes provided by the Official Statistical Institutes. Instead, the microdata are integrated, that is, they are transformed into a complex coding scheme which seeks to preserve all significant detail yet assign identical codes to identical concepts. The integrated microdata are provided only in the form of extracts, custom tailored to each researcher's needs. What this means is that there is no distribution of entire datasets in the form of of compact discs, DVDs or otherwise. Since each dataset is custom tailored, "collecting" or "boot-legging"datasets is not only illegal, but effectively curtailed. The database is so enormous and evolving so quickly that users and their institutions have a powerful interest in safe-guarding the data and promoting good use.

To request an extract, the user must first become licensed (see below) and then sign into the project website ("*create an extract*") by entering the registered password. Then a series of selections are made by means of point-and-click menus. The user selects the country or countries, census years, samples, and variables as well as the form of metadata required for the statistical package to be used (SAS, STATA or SPSS are provided). The IPUMS extract engine also makes it possible to select cases (persons, households, or dwellings) with specific characteristics, such as, say, females aged 15-19 in the workforce. Selected cases may also include members of households or families in which the selected case is found.

One of the most valuable enhancements of the database is the "SUBSAMPLE" feature. With SUBSAMPLE, the user may request any of 100 sub-samples each of which is nationally representative and preserves any stratification of the larger sample from which it is drawn. This tool may be used to test procedures, economize resources (where the research does not require large samples), or estimate variances through the replicate method.

Once the selections are complete, there is an opportunity to review or revise before final submission of the request. Then, once submitted, the extract engine registers the request and places it in a data processing queue. When the extract is ready (usually in a matter of minutes), the researcher is notified by email that the data should be retrieved within 72 hours. A link is provided to a password-protected site for downloading the specific extract. The data are encrypted during transmission using 128-bit SSL (Secure Sockets Layer) encryption standard, matching the level used by the banking and other industries where security and confidentiality are essential. The researcher may then securely download the file, decompress it and proceed with the analysis using the supplied integrated metadata consisting of variable names and labels. The metadata are in ASCII format so that a researcher may readily adapt them for use by any statistical software.

3. Confidentiality

IPUMS-International means Integrated Restricted-Access, Anonymized Microdata Extracts. The IPUMS-Europe acronym carries "PUMS" embedded in its name, but in fact the data are available only as "restricted-access extracts" from anonymized, integrated samples. Thus, "IRAAME" would be a more literal acronym, and indeed when the IPUMS was internationalized in 1998, the Principal Investigators discussed replacing "PUMS" with a more accurate moniker. We also discussed inserting "scientific" in place of "public". However, a decade-long, unbroken string of successes in securing monetary resources from the National Science Foundation and the National Institutes of Health dissuaded us then from abandoning the acronym, as it does now with the sister projects, IPUMS-Latin America and IPUMS-Europe.

Nonetheless, it is important to understand that a comprehensive array of additional protections, much greater than those for IPUMS-USA, are in place to guarantee the privacy and statistical confidentiality of census microdata samples incorporated into the IPUMS-International database. These protections involve three elements:

1. legal: dissemination agreements between the University of Minnesota and each participating Official Statistical Institute
2. administrative: licenses between the University of Minnesota and each user, specifying conditions and restrictions of use
3. technical: perturbations of the data (swapping, recoding, etc.) to make exceedingly unlikely the identification of individuals, families or other entities in the data. Technical measures have the additional benefit that any assertion of absolute certainty in identifying anyone in the data is false.

While much of the literature on statistical confidentiality ignores the legal and administrative environment (and in doing so exaggerates the risk of improper use), we remain firmly persuaded that the strongest system of protections must take into account all three types of guarantees (Thorogood 1999). IPUMS-International confidentiality standards seek to comply with EC Regulation 831/2002, although this regulation encompasses only four datasets at present: European Community Household Panel, Labor Force Survey, Community Innovation Survey, and Continuing Vocational Training Survey (King 2003).

3.1. Legal protections

First, with regard to legal protections, IPUMS-International projects are undertaken only in countries where explicit authorization is forthcoming, usually in the form of a memorandum of understanding endorsed by the official statistical institute and the legal authority of the University of Minnesota (see Appendix A). No work is begun with the microdata of a country without prior signed authorization from the corresponding OSI. The agreement is highly general and uniform across countries. Details specific to each country such as fees and sample densities are negotiated separately with each official agency and do not form part of the agreement. Under a carefully worded legal arrangement, the Regents of the University of Minnesota are responsible for enforcing the terms of the accords. The ten clauses spell out: 1) rights of ownership, 2) rights of use, 3) conditions of access (in which statistical institutes cede their gate-keeping authority to grant individual licenses to the IPUMS-International project), 4) restrictions of use, 5) the protection of confidentiality, 6) security of data, 7) citation of publications, 8) enforcement of violations, 9) sharing of integrated data, 10) and arbitration procedures for resolving disagreements. There are no secret clauses or special considerations. Although minor rewording of clauses is permissible, all members of the consortium are treated equally.

Nonetheless, the protocols are revised, indeed expanded, as OSIs suggest, or request, modifications. Any request for modification is reviewed by the legal cabinet of the University of Minnesota. Compare for example the violations clause in Appendix A (as signed by Statistics Austria in January 2002) with the current text (additions in italics), as follows:

> 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 [the Statistical Agency of Country X] will assist in the enforcement of provisions of this accord.*

Recently the tenth clause, which establishes jurisdiction for the settlement of a dispute between the University and any signatories to the memorandum, was amended, substituting the International Court of Arbitration for the Chamber of Commerce of Paris. At the same time, an eleventh clause, regarding order of precedence, was added, specifying that the clauses in the letter of understanding supersede any contract, purchase order or other document signed between the parties. Under the agreement, the Minnesota Population Center and its authorized partners are obliged to share the integrated data and documentation with the official statistical institutes and to police compliance by users.

3.2. Administrative measures

Second, researchers must apply for a license to gain access to the microdata extraction system (see: https://www.ipums.org/international/apply_for_access.html). Grounds for approval are based upon three considerations:

1. whether the data are appropriate for the proposed project as stated in the applicant's project description
2. whether the applicant is an academic, non-commercial user
3. whether the applicant agrees to abide by the restrictions on conditions of use (see Appendix B).

The vetting of applications is performed by the Principal Investigators of the IPUMS-International project. It is noteworthy that approximately one-third of applications are denied because of a failure to adequately satisfy one or more of the specified conditions. It is gratifying to report that few users appeal denial of access.

Administrative measures limit access to the extract system to users, who:

1. sign the electronic non-disclosure license;
2. endorse prohibitions against a) attempting to identify individuals or the making of any claim to that effect, b) reporting statistics that might reveal an identity and c) redistributing data to third parties;
3. agree to use the data solely for non-commercial ends and to provide copies of publications to ensure compliance;
4. place themselves under the authority of educational institutions, employers, institutional review boards, professional associations, and other enforcement agencies to deal with any alleged violations of the license.

The license is granted to users, individually, not to research groups, classes, or institutions. The license application instructs the applicant regarding conditions of use (see Appendix B). The license is not transferable. Should the individual change institutions or employment, the license must be updated. Data can be reassigned within an institution, but the person responsible for the microdata must apply for access. Once licensed, the user is permitted to download data extracts of samples and variables according to need. Licensees import the extracts into their statistical software of choice to analyze at the convenience in their own institutional setting.

Since its adoption in 2002, the basic application procedure remains unchanged. Few suggestions for enhancing the application form or approval process have been forthcoming, even though advice is solicited from users, statistical institutes, funding agency review boards, and outside experts. Nevertheless in 2006, we plan to strengthen application and vetting procedures, primarily to guard against fraudulent applications. In addition to requesting additional details about the applicant and institutional affiliation, the form will contain the following statement as a heading:

> **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. Violation of this agreement may lead to professional censure, loss of employment, 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.**

In the United State, an Institutional Review Board for the protection of human subjects is required of any academic research institution applying for funding from the National Institutes of Health. IRBs provide a strong mechanism for enforcing of the IPUMS-International license agreement in the United States. Most developed and developing countries have similar mechanisms. Delegates to this conference are invited to provide the names of similar institutions in their country. Oversight boards are nearly universal. It is these boards that provide a strong shield for insuring the highest standards of scientific conduct.

Finally, once these revisions to the application are in place on the website, licenses will be valid for one year and will be renewable. A license may be suspended at any time.

3.3. Technical protections

Third are the technical measures taken to ensure statistical confidentiality. Sampling of datasets alone "provides the additional uncertainty needed to protect many data releases…" (Anderson and Fienberg 2001). Census errors and non-response error also provide their own confidentiality protections.
As Fienberg (2005) has noted the principal threats are geographic detail and extreme values. Many statistical institute partners anonymize the microdata and implement technical measures of confidentiality protection before the data are entrusted to the project. When the OSI provides a sample that is also made available to others–such as public use samples, SARs and the like–no additional protections are implemented by the project. Usually the project is not informed of the precise technical measures imposed on the data. Where the samples are unique, we impose the following technical protections (based on Thorogood 1999):

1. adopt sample density according to official norms or conventions (see tables 1 and 2);
2. limit geographical detail by means of global recoding to administrative units with a minimum number of inhabitants. For some countries, this limit is as high as 100,000 and for others as low as 10,000. For the European project, NUTS3 is likely to be the lowest level of identifiable administrative geography, with the minimum threshold varying from 20,000 to 100,000 inhabitants according to the most recent census;
3. top and bottom code unique categories of sensitive variables (identified by the OSI);
4. round, group, or band age as necessary;
5. suppress date of birth (only age is provided);
6. suppress detailed place of birth (<20/100,000 population);
7. suppress detailed place of residence, work, study, and migration (<20/100,000 population);
8. systematically "swap" (recode) place of enumeration for a fraction of households, inversely proportional to population size at the NUTS3 level; Data swapping protects confidentiality by introducing uncertainty about sensitive data values, yet maintains the strength of statistical inferences by preserving summary statistics (see Fienberg and McIntyre, 2004).
9. randomly order households within administrative units (NUTS3);
10. and, conduct a sensitivity analysis once these measures are imposed to determine what additional measures may be required.

We continue to evaluate emerging methods and technologies for disclosure protection (McCaa and Ruggles 2002). At present we have decided against automatic data protection methods such as *μ-Argus* (Hundepool et al, 1998; Polettini and Seri 2003). It should also be noted that no synthetic data are added to the IPUMS samples.

4. Shifting the R-U Curve Rightward

In practice, disclosure of confidential information from census microdata samples is highly improbable. Moreover, researchers have no interest or incentive to even attempt to identify individuals. There are compelling reasons for jealously guarding confidentiality, both for individual users and the academic community as a whole. Any partially successful effort, such as that by a rogue intruder, will require an enormous investment of resources to obtain rather trivial details invariably with a high degree of uncertainty about whether any one record truly corresponds to a targeted individual or entity

(Dale and Elliot 2001). Indeed, over the past forty years of disseminating census microdata in the United States and elsewhere there are few allegations of misuse or breach of statistical confidentiality by an academic researcher. The IPUMS-International procedures are designed to extend this nearly perfect record.

Len Cook (2003) notes that increased access is not a threat to statistical systems. On the contrary he observes that increasingly there is an expectation that analysis of microdata will inform research and evaluation of policy. Increased access builds trust in statistical systems, while lack of access leads to suspicion. He advocates that different forms of access be granted for different degrees of trust. Moreover academic researchers possess a range and depth of expertise that national statistical institutes cannot replicate.

Julia Lane (2003) highlights five classes of benefits which accrue from broader access to microdata: address more complex questions, calculate marginal effects, replicate findings, assess data quality and build new constituencies or stakeholders. Replication is extremely important because there is an overwhelming temptation for scientists to misrepresent results when the data are unlikely to be available to others. The IPUMS system facilitates replication by providing access to microdata to all approved academic users on an equal basis.

5. Conclusion

Now that the construction of anonymized microdata data samples is becoming an increasingly widespread practice, harmonization of census microdata is an obvious next step to enhancing use. With the emergence of global standards of statistical confidentiality and the massive power of ordinary desktop computers, the major challenge that remains is the actual construction of integrated, anonymized census microdata samples. By restricting access to a class of academic users, high-density microdata extracts can be provided to researchers at vanishingly low risk.

References

Anderson, Margo and Stephen E. Fienberg. (2001). "U.S. Census Confidentiality: Perception and Reality," International Statistical Institute Biennial Meeting (Seoul).

Cook, Len. (2003). "Summary of Discussants' Main Points," in *Statistical Confidentiality and Access to Microdata: Proceedings of the Seminar Session of the 2003 Conference of European Statisticians*. Geneva, pp. 7-10.

Dale, A. and Elliot, M. (2001). 'Proposals for 2001 SARS: An assessment of disclosure risk.' *Journal of the Royal Statistical Society, Series A,* 164, part 3, pp.427-447.

Fienberg, Stephen E. (2005). "Confidentiality and Disclosure Limitation," *Encyclopedia of Social Measurement*, Elsevier, Inc.. Vol. 1, pp. 463-469.

Hundepool, A., L. Willenborg, A. Wessels, L. van Gemerden, S. Tiourine and C. Hurkens. (1998). *µ-Argus User's Manual*. Statistics Netherlands: Voorburg.

King, John (2003). "Recent European Union Legislation for Research Access to Confidential Data: Implementation and Implications," in *Statistical Confidentiality and Access to Microdata: Proceedings of the Seminar Session of the 2003 Conference of European Statisticians*. Geneva, pp. 97-116.

Lane, Julia (2003). "Uses of Microdata: Keynote Speech," in *Statistical Confidentiality and Access to Microdata: Proceedings of the Seminar Session of the 2003 Conference of European Statisticians*. Geneva, pp. 11- 20.

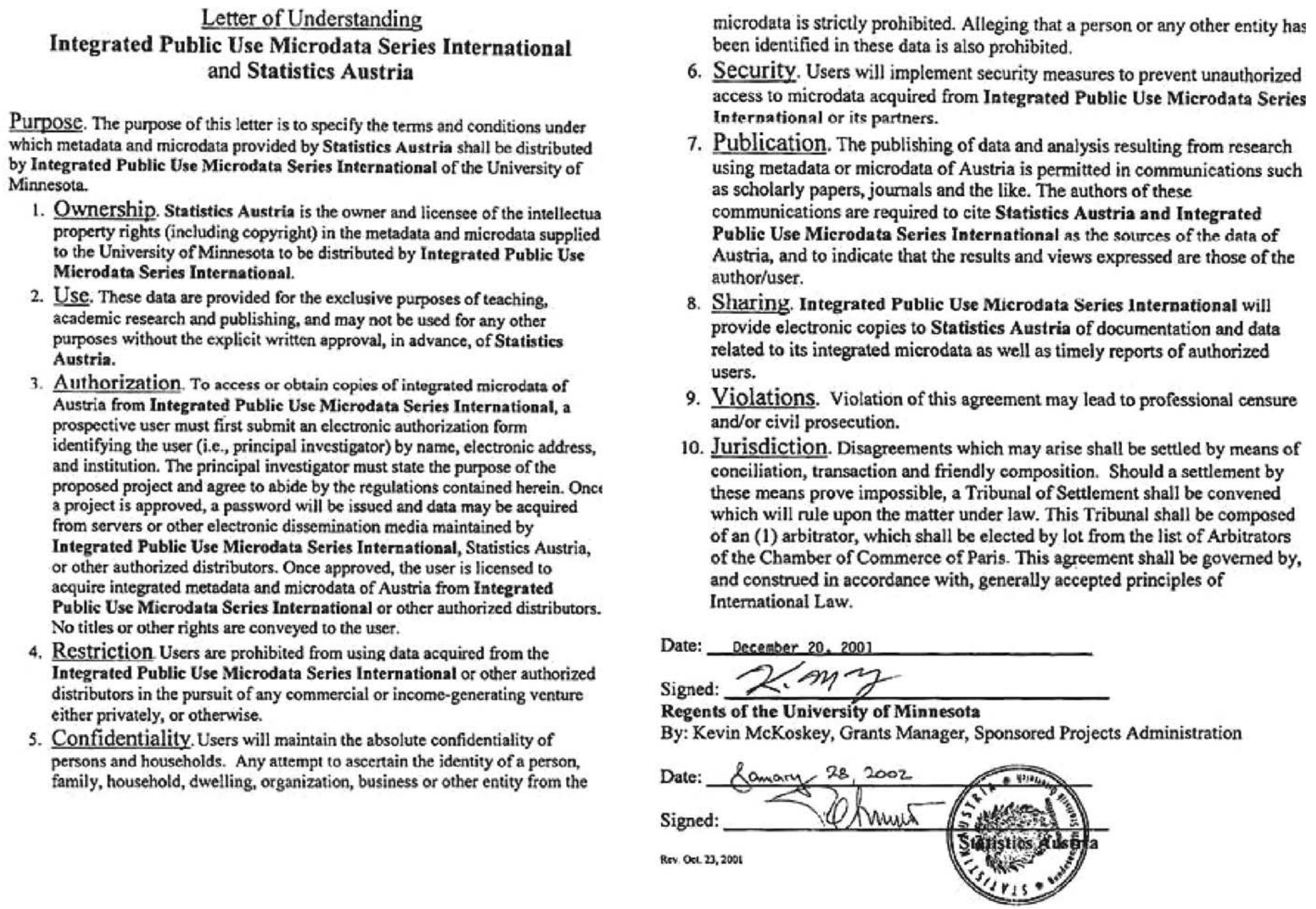
McCaa, Robert and Albert Esteve. (2005). "La integración de los microdatos censales de América Latina: el proyecto IPUMS," *Estudios Demográficos y Urbanos* 20:1(58) 37-70.

McCaa, Robert, and Steven Ruggles. (2002). The Census in Global Perspective and the Coming Microdata Revolution. In Vol. 13, *Nordic Demography: Trends and Differentials, Scandinavian Population Studies*, edited by J. Carling. Oslo: Unipub/Nordic Demographic Society, pp. 7-30.

Polettini, Silvia and Giovanni Seri (2003). “Guidelines for the Protection of Social Microdata Using Individual Risk Methodology: Application within µ-Argus Version 3.2,” in *CASC Project: Computational Aspects of Statistical Confidentiality*.

Ruggles, Steven, Miriam King, Deborah Levison, Robert McCaa, and Matthew Sobek. (2003). “IPUMS-International: An Overview”. *Historical Methods*, 36: 60-65.

Thorogood, D. (1999). ‘Statistical Confidentiality at the European Level.’ Paper presented at: Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Thessaloniki, March.



Letter of Understanding

Integrated Public Use Microdata Series International and Statistics Austria

Purpose. The purpose of this letter is to specify the terms and conditions under which metadata and microdata provided by **Statistics Austria** shall be distributed by **Integrated Public Use Microdata Series International** of the University of Minnesota.

1. Ownership. **Statistics Austria** is the owner and licensee of the intellectua property rights (including copyright) in the metadata and microdata supplied to the University of Minnesota to be distributed by **Integrated Public Use Microdata Series International.**
2. Use. These data are provided for the exclusive purposes of teaching, academic research and publishing, and may not be used for any other purposes without the explicit written approval, in advance, of **Statistics Austria.**
3. Authorization. To access or obtain copies of integrated microdata of Austria 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. Onc 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,** Statistics Austria, or other authorized distributors. Once approved, the user is licensed to acquire integrated metadata and microdata of Austria 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 Austria is permitted in communications such as scholarly papers, journals and the like. The authors of these communications are required to cite **Statistics Austria and Integrated Public Use Microdata Series International** as the sources of the data of Austria, and to indicate that the results and views expressed are those of the author/user.
8. Sharing. **Integrated Public Use Microdata Series International** will provide electronic copies to **Statistics Austria** of documentation and data related to its integrated microdata as well as timely reports of authorized users.
9. Violations. Violation of this agreement may lead to professional censure and/or civil prosecution.
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 (1) arbitrator, which shall be elected by lot from the list of Arbitrators of the Chamber of Commerce of Paris. This agreement shall be governed by, and construed in accordance with, generally accepted principles of International Law.

Date: December 20, 2001

Signed:

Regents of the University of Minnesota

By: Kevin McKoskey, Grants Manager, Sponsored Projects Administration

Date: January 28, 2002

Signed:

Statistics Austria

STATISTIK AUSTRIA

Rev. Oct. 23, 2001