

United Nations Economic Commission for Europe  
Conference of European Statisticians



# Managing Statistical Confidentiality & Microdata Access

Principles and  
Guidelines of  
Good Practice



U N I T E D N A T I O N S

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](mailto: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 Brügger (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



## I. INTRODUCTION

1. Historically, confidentiality protection has been mainly a national issue. However, in the context of increasing data dissemination over the Internet, it is now also becoming an international issue. There is a great deal of international collaboration among members of the research community, and the researchers can be very critical towards different access rules in different countries. Furthermore, researchers are often not allowed to access other countries' microdata for fear that confidentiality protection cannot be guaranteed. Nevertheless, cross-country comparisons can be a very important part of a research project. This is not only of interest to academic researchers. International agencies are among those who want to use microdata for research purposes, particularly cross-country comparisons. Such studies are usually of great interest and relevance to the participating countries.
2. This raises the question of whether it is possible to internationally agree on some common principles for dissemination of microdata. This question should be seen in the context of the 2003 Conference of European Statisticians (CES) agreeing that support for research is an important activity of the National Statistical Offices (NSOs), and generally NSOs could do more to satisfy these needs. Doing more includes providing access to microdata which is the main focus of these principles and guidelines. (Although the reference is to NSOs in these guidelines, in many countries, particularly those with decentralised systems, there are several statistical producers. The reference to NSOs should be read as incorporating all producers of official statistics.)
3. There are two key objectives in these guidelines:
  - (i) to foster greater uniformity of approach by countries whilst facilitating better access to microdata by the research community for worthwhile purposes; and
  - (ii) through these guidelines and supporting case studies, to enable countries to improve their arrangements for providing access to microdata.
4. The term microdata is used throughout the paper. It can refer to data about an individual person, household, business or other entity. It may be data directly collected by the NSO or obtained from other sources, such as administrative sources.
5. These guidelines recognise that the precise arrangements for access to microdata will vary from country to country. They will depend on matters such as legislation, public attitudes and the capacity to support the research community. For example, the arrangements for a well-developed statistical office will be quite different from those in a less well-developed statistical office. Some countries may also feel they do not have either the systems or resources to maintain the necessary confidentiality safeguards. It should not be anticipated that each country will come up with precisely the same arrangements, although it is hoped these guidelines will lead to greater uniformity of approach.
6. We should also be mindful that not all countries are coming from the same position. Some countries, particularly from Eastern Europe, have traditionally not had strong legislation supporting confidentiality. This is being changed in many cases but the cultural change to support the legislative change can take longer.

7. A number of countries have existing legislation. Also the European Union (EU) has legislation on confidentiality that embodies several principles and rules. These will already be applied by many ECE countries, especially the EU countries. It is recognised that existing legislation is not easily changed and that changes to existing guidelines require collaboration with a range of stakeholders. But opportunities do arise from time to time and these guidelines and the associated case studies may be useful in determining appropriate changes. Indeed, in some countries, these guidelines may provide a useful stimulus for debating and agreeing on changes.

8. Any questions on these guidelines should be submitted by e-mail to the Statistics Division of the UNECE at [confidentiality@unece.org](mailto:confidentiality@unece.org).

## II. WHY SHOULD NATIONAL STATISTICAL OFFICES SUPPORT THE RESEARCH COMMUNITY?

9. In most countries, official statistics are collected not just for governments but for the use of the community. This is particularly the case in democracies where official statistics can be used to assess the effectiveness of governments' policies and programmes – they provide a mirror on society.

10. To quote a 1993 White Paper on Open Government in the United Kingdom:

“Official statistics are collected by government to inform debate, decision making and research both within government and by the wider community.

“They provide an objective perspective of the changes taking place in national life and allow comparisons between periods of time and geographical areas.

“Open access to official statistics provides the citizen with more than a picture of society. It offers a window on the work and performance of government itself, showing the scale of government activity in every area of public policy and allowing the impact of public policies and actions to be assessed.”

11. The research community plays a particularly important role in stimulating policy analysis and debate and assessing the effectiveness of government programmes. This requires access to good-quality statistical data if their analyses are to be effective. If they do not have access to relevant official statistical data, they will often seek to collect their own data. As well as incurring additional costs to both the data collector and the respondent, these collections will often be of lower quality.

12. Providing researcher access to microdata can also be a way of extracting additional value from the cost of collecting official statistics, and of obtaining valuable insights into the quality of the data and how statistical surveys might be improved or extended.

13. What is the research community? It includes those working in academic institutions, of course. It also includes researchers working in non-government organizations and international agencies. Furthermore, some researchers requiring access to microdata will work within government-funded agencies and institutions. For the purposes of these Guidelines, all of these researchers are regarded as part of the “research community”. However, as will be seen from the Guidelines, the pertinent issues may vary somewhat between the different elements of the research community.

14. The following sections try to bring together the perspectives of national statistical offices and the research community with the intention of trying to find arrangements that largely satisfy the needs of both groups. These are considered in more detail in Chapter 6. The perspective of the National Statistical Office.

15. NSOs must maintain the trust of respondents if they are to continue to cooperate in their data collections. Confidentiality protection is the key element of that trust. If respondents believe or perceive that a NSO will not protect the confidentiality of their data, they are less likely to cooperate or provide accurate data. One incident, particularly if it receives strong media attention, could have a significant impact on respondent cooperation and therefore on the quality of official statistics.

16. This is the dominant issue from the point of view of NSOs but there are other concerns. A key one is whether they have sufficient authority to support researcher access to microdata, either through a legal mandate or some other form of authorisation.

17. Some NSOs are concerned that the quality of their microdata may not be good enough for further dissemination. Whilst quality may be sufficiently accurate to support aggregate statistics, this may not be the case for very detailed analysis. In some cases, adjustments are made to aggregate statistics at the output editing stage without amendment to the microdata. Consequently, there may be inconsistencies between research results based on microdata and published aggregate data.

18. NSOs may also be concerned about costs. These include not only the costs of creating and documenting microdata files, but the costs of creating access tools and safeguards, and of supporting and authorising enquiries made by the research community; new users of data files need help to navigate complex file structures and variable definitions. Although the costs are borne by the NSOs, they are usually not provided with budget supplementation to do the additional work. And on the whole, researchers do not have the funding to contribute substantially to these costs.

19. On the other hand, NSOs are increasingly recognising the importance of supporting the research community, and of the additional value that is provided to NSO data collection and processing effort through effective use of its data for research. Specifically, it is in the public interest that insights, which can be provided from the data, can be made available to decision makers and the public. Furthermore, if survey data are used more extensively in this way, it can provide an extra level of protection against budget reductions to these statistical programmes.

### **The perspective of the research community**

20. From the perspective of the research community, supporting research based on microdata should be an important component of any official statistical system. The benefits include the following:

- (i) microdata permits policy makers to pose and analyse complex questions. In economics, for example, analysis of aggregate statistics does not give a sufficiently accurate view of the functioning of the economy to allow analysis of the components of productivity growth;
- (ii) access to microdata permits analysts to calculate marginal rather than just average effects. For example, microdata enable analysts to do multivariate regressions whereby the marginal impact of specific variables can be isolated;
- (iii) broadly speaking, widely available access to microdata enables replication of important research;

- (iv) access to microdata for research purposes, and the resulting feedback, can facilitate improvements in data quality. For example, the US Bureau of the Census has formalised the documentation it requires from researchers to assist it in improving the quality of its surveys;
- (v) it increases the range of outputs derived from statistical collections and hence the overall value for money obtained from these collections.

21. Furthermore, lack of access to microdata may result in researchers developing and conducting their own statistical collections, adding to the reporting burden imposed on the community. As well as the cost involved (to the collector as well as the respondents), the collections will usually be of inferior quality and with smaller samples than official surveys. This will lead to lower quality research results. There are benefits from having an accepted and authoritative, as well as high quality, data source for all analysis compared with the alternative of researchers using different data sets to analyse particular topics. NSOs can play a very useful role in this respect.

22. The researchers point out that they are not interested in identifying individuals and the evidence is that this is indeed the case. Given this, they feel that NSOs have generally been too conservative in the access they provide to microdata.

23. At a 2003 Workshop on Confidentiality Research hosted by the United States National Science Foundation, Peter Madsen referred to the Privacy Paradox. He argues that “the rush to ensure complete levels of privacy in the research context paradoxically results in less social benefit, rather than in more”. He argues that when you include the concept of utility you may get different outcomes:

“Perhaps through this additional concept of utility, people will recognise that while they surely have the right to privacy, they may also come to the realisation that they have a duty to share information, if the common good is to be furthered.”

Some use the term “privacy deficit” recognising that there are privacy issues associated with microdata release. The discussion can then focus on whether the benefits of a proposal outweigh any privacy deficit.

24. The research community also sees the importance of research into improved methods of confidentiality protection that increase the usefulness of the underlying data. NSOs would agree with the importance of this research. However, this research is only likely to lead to a partial answer to the desire for improved access to microdata for research purposes and researchers would remain frustrated if NSOs relied solely on improved statistical methods for confidentiality protection.



## CORE PRINCIPLES

25. The sixth United Nations Fundamental Principle of Official Statistics is very clear on statistical confidentiality.

“Individual data collected by statistical agencies for statistical compilation, whether or not they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.”

Any principles for microdata access must be consistent with this Fundamental Principle.

26. The following principles should be used for managing the confidentiality of microdata. Each is discussed in the following paragraphs.

Principle 1: It is appropriate for microdata collected for official statistical purposes to be used for statistical analysis to support research as long as confidentiality is protected.

Principle 2: Microdata should only be made available for statistical purposes.

Principle 3: Provision of microdata should be consistent with legal and other necessary arrangements that ensure that confidentiality of the released microdata is protected.

Principle 4: The procedures for researcher access to microdata, as well as the uses and users of microdata, should be transparent and publicly available.

27. Making microdata available for research is not in contradiction with the sixth UN Fundamental Principle as long as it is not possible to identify data referring to an individual. Principle 1 does not constitute an obligation to provide microdata. The National Statistical Office should be the one to decide whether to provide microdata or not. There may be other concerns (for example, quality) that make it inappropriate to provide access to microdata. Or there may be specific persons or institutions to whom it would be inappropriate to provide microdata.

28. For Principle 2, a distinction has to be made between statistical or analytical uses and administrative uses. In the case of statistical or analytical use, the aim is to derive statistics that refer to a group (be it of persons or legal entities). In the case of administrative use, the aim is to derive information about a particular person or legal entity to make a decision that may bring benefit or harm to the individual. For example, some requests for data may be legal (a court order) but inconsistent with this principle. It is in the interest of public confidence in the official statistical system that these requests are refused. If the use of the microdata is incompatible with statistical or analytical purposes, then microdata access should not be provided. Ethics committees or a similar arrangement may assist in situations where there is uncertainty whether to provide access or not.

29. Researchers are accessing microdata for research purposes but to support this research they may need to compile statistical aggregations of various forms, compile statistical distributions, fit statistical models, or analyse statistical differences between sub-populations. These uses would be consistent with statistical purposes. To the extent that this is how the microdata are being used, it could also be said to support research purposes.

30. With respect to Principle 3, legal arrangements to protect confidentiality should be in place before any microdata are released. However, the legal arrangements have to be complemented with administrative and technical measures to regulate the access to microdata and to ensure that individual data cannot be disclosed. The existence and visibility of such arrangements (whether in law or supplementary regulations, ordinances, etc.) are necessary to increase public confidence that microdata will be used appropriately. Legal arrangements are clearly preferable but in some countries this may not be possible and some other form of administrative arrangement should be put in place. The legal (or other arrangements) should also be cleared with the privacy authorities of countries where they exist before they are established by law. If such authorities do not exist, there may be NGOs who have a “watchdog” role on privacy matters. It would be sensible to get their support for any legal or other arrangements, or at least to address any serious concerns they might have.

31. In some countries, authorising legislation does not exist. At a minimum, release of microdata should be supported by some form of authority. However, an authorising legislation is a preferable approach.

32. Principle 4 is important to increase public confidence that microdata are being used appropriately and to show that decisions about microdata release are taken on an objective basis. It is up to the NSO to decide whether, how and to whom microdata can be released. But their decisions should be transparent. The NSO web site is an effective way of ensuring compliance and also for providing information on how to access research reports based on released microdata.

## IV. SUPPORTING LEGISLATION

33. Legislation supporting microdata release is very important as highlighted by Principle 3 (see Chapter 3). There are several reasons:

- (i) to provide public confidence in the arrangements – that there are legal constraints that determine what can and cannot be done;
- (ii) to provide mutual understanding between NSOs and researchers on the arrangements;
- (iii) to provide for greater consistency in the way research proposals are treated; and
- (iv) to provide a basis for dealing with breaches.

34. The legislation need not exist in primary legislation or law. The detail may be better suited to regulations, ordinances, etc. that still have some legal impact. If legislation is not available, some other form of authorisation is essential. The reputation of the NSO is at risk if there is not some form of authority to enable the release of microdata even when anonymised.

35. It is important that the legislation (or authorisation) covers the following aspects:

- (i) what can and cannot be done and for what purposes;
- (ii) the conditions of release; and
- (iii) the consequences if these conditions are breached.

36. Case Studies 1 and 2 outline the legislation for the provision of microdata for Australia and Finland respectively.

## V. METHODS OF SUPPORTING THE RESEARCH COMMUNITY

37. There are various ways a National Statistical Office (NSO) can support research work. These are summarised below. There is more expansive commentary in the following sections. Case studies are used to further illustrate these different methods.

38. There is one important point which is not always understood. Microdata files can be anonymised by removing names and addresses and taking other steps (e.g. collapsing geographical detail) to ensure that the identification of individuals is highly unlikely when these files are looked at in isolation. This could be referred to as eliminating spontaneous identification. But other microdata files exist in the public and private sector, sometimes with individuals identified. Studies have shown that by statistically matching the NSO microdata files with existing files, unique records can be identified. The number can be quite significant depending on the amount of detail available in the NSO microdata file. Also in relative terms, the number of unique cases will be greater for smaller countries. These risks are not always well understood by NSOs. Of course, they will be reduced if techniques such as data perturbation or data swapping are used in the NSO microdata file.

- (i) Statistical products for use outside the NSO:

Dissemination Stream	Notes
Statistical Tables and Data Cubes	This can include both standard tables and special tables (or special analyses for that matter) generated at the request of the researcher. Some offices now release very detailed matrices, known as data cubes, which researchers can manipulate to support their own needs. However, if these are very detailed, the level of confidentiality risk can be similar to microdata.
Anonymised Microdata Files (AMFs) <ul style="list-style-type: none"> <li>- Public Use Files (PUFs)</li> </ul>	These are microdata files that are disseminated for general public use outside the NSO. They have been anonymised and are often released on a medium such as CD-ROM, sometimes through a data archive. (Note: The term anonymised implies that not only are names and addresses removed, but other steps are taken (e.g. collapsing of geographic details) to ensure that identification of individuals is highly unlikely.) The level of confidentiality protection in Public Use Files should be such that identification is not possible even when matched with other data files. Public Use Files are a way of providing access to researchers in some countries.

<p>Anonymised Microdata Files - licensed files</p>	<p>Licensed files are also anonymised but are distinct from Public Use Files in that their use is restricted to approved researchers and an undertaking or contract is signed before files are provided to the researchers. Even if advertised as generally available to the public, they are not released before an undertaking or contract is provided by the researcher. Even though anonymised and other steps are taken to ensure that identification of individuals is highly unlikely when used in isolation, they may contain potentially identifiable data if linked with other data files; this is one reason why a preventive undertaking or contract is required. There may be other conditions of use that the NSO may impose on researchers.</p>
--	--

(ii) A service window through which researchers can submit data requests

Service	Notes
<p>Remote Access Facilities (RAFs)</p>	<p>Arrangements are now being made in many countries that allow researchers to produce statistical outputs from microdata files through computer networks, without the researchers actually ‘seeing’ the microdata. Because of the additional controls that are available through RAF, and the fact that microdata do not actually leave the NSO, access to more detailed microdata can be provided this way.</p>

(iii) Arrangements for allowing researchers to work on the premises of the National Statistical Office

Service	Notes
<p>Data Laboratories (DL)</p>	<p>On-site access to more identifiable microdata, typically with stringent audit trails and NSO supervision. The access to more detailed data creates some inconvenience to the researcher, because of the requirement of working at the NSO, or at an NSO enclave.</p>

### Statistical tables and data cubes

39. Statistical tables remain the most economical way of satisfying many research needs. Their importance should not be underestimated. The advent of data cubes (very detailed multi-dimensional tables) has increased the usefulness of statistical tables for research purposes as they allow researchers to manipulate the data cubes to suit their own needs.

40. Statistics Netherlands was one of the early organizations to embrace data cubes. Case Study 3 illustrates how they use data cubes as a key part of their dissemination strategy.

41. Confidentiality issues still exist for statistical tables and data cubes. For example, most statistical legislation requires that identifiable data cannot be released through statistical tables. But the ‘confidentialisation’ is done prior to release. Software systems exist for confidentialising statistical tables and improved methods continue to be developed. They are often referred to as disclosure avoidance methods.

### **Anonymised microdata files – Public Use Files**

42. This is seen as a very valuable service by researchers. However, in light of the increased possibilities for data matching, the trend might be to reduce the amount of data available in public use files and to put more reliance on licensed anonymised microdata files, RAFs and data laboratories for researcher access. In addition to steps taken to reduce identification, licensed files rely on researchers honouring the undertakings or contracts that they make not to attempt to identify. Such undertakings are often a key part of a release of licensed AMFs (see next section).

43. Although NSOs generally provide equality of access to users of their statistics, this may not be appropriate for microdata. A different attitude may be taken to users who do not have strong bona-fide research credentials or if they have access to databases where it would be easy to match AMFs.

44. The exception is Public Use Files (PUFs) where access is deliberately intended to be broad. Researchers have emphasised the importance of PUFs. They are greatly appreciated in those countries where they exist and they are used extensively. Yet it may not be difficult for someone who is so inclined to publicly identify some individuals through statistical matching with other databases, particularly for countries with smaller populations and those with population registers. Prior to the release of PUFs, there should be a close examination of the conditions under which they are released to better manage the risks of a confidentiality violation. For example, a legally enforceable agreement may be one of the requirements of access. It should be possible to set up an arrangement where a prior agreement needs to be signed even where access to PUFs is through the Internet. Generally, the level of risk will be much greater for countries with smaller populations. Consequently, researchers should not expect that all countries will release PUFs.

45. The risk of identification can be reduced by the use of techniques such as data swapping and data perturbation. These techniques are frequently used in the United States for example. The downside is that these techniques may reduce the usefulness of the underlying microdata.

46. Case Study 4 describes the arrangements for the release of PUFs in the United States. It is interesting to note the role that Social Data Archives play in managing access to PUFs to individual researchers.

47. There is extensive literature available on the methods for anonymising microdata files. A good summary is available in Willenburg, L. & de Waal, T. (2001), *Elements of Statistical Disclosure Control*. The software package,  $\mu$ -ARGUS, is concerned with protection of microdata against disclosure. Several techniques are available in  $\mu$ -ARGUS.

### **Anonymised microdata files - licensed files**

48. This is an arrangement where specific users are authorised or licensed to use anonymised microdata files after making a relevant undertaking or contract. Although these files have been anonymised and individuals cannot be identified from these microdata files in isolation, it may be possible to do so by (statistical) matching with other files, hence the need for a licence. There will be conditions associated with the licence, which can be specified in the undertaking or contract signed by the researcher or their institution. The conditions may vary from country to country or even from one researcher to another depending on the research proposal and possibly the affiliation of the researcher.

49. The conditions may include some or all of the following:

- an agreement by the researcher that he or she will abide by the conditions of release;
- that no attempt will be made to identify particular persons or organizations;
- the information will only be used for statistical or research purposes;
- the microdata will not be provided to other persons;
- the microdata will be returned to the NSO when the research project is completed; and
- no attempt will be made to statistically match with other databases without permission.

50. It is good practice for such an undertaking to have some legal standing, for example by providing for such undertakings within enabling legislation. This would allow legal actions to be taken in respect of breaches of the conditions of the undertaking. This does not preclude other actions that might be taken in respect of breaches such as not providing any further services to the researcher and/or possibly the researcher's institution. These are discussed in Chapter 7.

51. It should be possible to release more data through licensed files than public use files if reliance can also be put on the undertaking to ensure protection of the confidentiality of the data. That is, in cases where some of the data are potentially identifiable, when linked with other files.

52. Case Studies 5, 6 and 7 describe the arrangements for the release of licensed microdata files in Australia, Netherlands and Sweden respectively.

### **Microdata where identification may be possible**

53. Some countries externally release microdata files for statistical or research purposes containing data which might be identifiable albeit under strict licensing agreements. The licensing agreements should include the conditions under which the data can be used, and the procedure should be specifically covered by legislation. A strict procedure is necessary in order to maintain respondents' confidence and the general public's trust. Remote access facilities and data laboratories are other ways of dealing with this type of situation.

### Remote Access Facilities (RAFs)

54. These facilities are becoming increasingly important but the way RAFs are implemented varies considerably from country to country. The key characteristic is that researchers do not have access to the microdata itself but tasks using that microdata can be submitted remotely over the Internet. Often there is a contractual arrangement between the NSO and the researcher or the institution of the researcher.

55. By way of illustration, Statistics Canada provides researchers with dummy microdata files and allows them to submit runs against the full file via computer networks. Statistics Canada runs the requests offline and sends the results back via computer networks after checking for confidentiality. Although similar arrangements exist at the Australian Bureau of Statistics, there are some important differences. The microdata files are confidentialised to prevent spontaneous identification before becoming accessible through a RAF. However, trial runs are permitted against the RAF files and small numbers of unidentified unit records are allowed to be downloaded to explore outliers and the like. Output is checked before being sent to the researcher. The system currently operates in batch mode but an interactive version is being considered. The arrangements in Statistics Denmark are different again. It is an on-line system where researchers can run analyses against the full microdata file. Arrangements are such that they cannot download the microdata itself. To further manage risks, they rely on the agreements made by institutions and the retribution (particularly denial of future access) if there are breaches of the agreements.

56. There are two basic types of RAF.

- (a) Remote execution, where a researcher submits a program and receives the output later over the Internet.
- (b) Remote facilities, where the researcher performs the analysis and can immediately see the answer on the screen.

Many countries have facilities along the lines of (a) but, apart from the Danish system, facilities along the lines of (b) are still being developed. The acceptability of different arrangements is likely to vary country by country.

57. Although only available so far in a few countries, and though the models and approaches vary as illustrated above, the experience to date with remote access facilities has generally been positive.

58. From the cost perspective, RAFs are preferable to data laboratories (see below) as the supervised access in a RAF is less labour intensive than the supervised use involved in data laboratories.

59. If these facilities do not remove identification risk entirely, there should still be some agreement made by researchers to ensure they are fully aware of their obligations. It is good practice to only provide access to those researchers who have signed some form of agreement outlining the conditions of access. Education of the RAF users is also important, together with regular monitoring and checking of the use of these facilities.

60. Case Studies 8, 9 and 10 outline the remote access facilities in place in Canada, Australia and Denmark respectively.

### **Data laboratories**

61. Data laboratories have been in use for many years in some NSOs and have been effective in controlling identification risk whilst enabling researcher access, particularly for data sets where release of a confidentialised microdata file is not possible. They still require conditions of access to provide an adequate level of protection. The main criticism of DLs has been the lack of convenience to the researcher, including sometimes being forced to use unfamiliar data analysis software. They are also expensive for the NSO to manage compared with other options.

62. Some NSOs (e.g. Statistics Canada) have established new premises for data laboratories in locations that are more convenient to researchers (sometimes known as Research Data Centres), but this can also be an expensive option unless specific funding is provided to the NSO.

63. What are key conditions of access to microdata through data laboratories? These might include (a) documentation of the public good that the research will provide, (b) outlining how the results will be accessible to the public, (c) evidence of the bona fides of researchers, (d) a legally binding undertaking, and (e) requirements for supervision by NSO staff.

64. Case Studies 11, 12, 13, 14, 15 and 16 outline the data laboratory arrangements in Canada, USA, Netherlands, New Zealand, Italy and Brazil respectively.

### **Engaging a researcher as a temporary NSO staff member**

65. Another way that researchers may access microdata is through their being engaged as temporary NSO staff members and making them subject to the same secrecy provisions as the staff of the NSO. This should not be done unless the researcher is assisting with the work of the NSO, otherwise the arrangement could be seen as a sham. If this type of pretence were occurring and became public, confidence in the NSO would diminish.

66. The involvement of the researcher may be at the initiation of the NSO, if the researcher is seen as someone who can bring special skills to the work of the NSO and extend the usefulness of the data set. On the other hand, the proposal may come at the initiation of the researcher. It is easier to demonstrate that researchers are assisting the NSO if a published NSO output will result from the work (even if branded somewhat differently from normal published outputs). Of course, there will be benefits for the researchers from such arrangements and it may be agreed that they might publish the outputs of their research in other ways (perhaps after clearance by the NSO).

## Business data

67. There are some special issues associated with business data, including agricultural businesses. Businesses, and in particular large businesses, are more easily identifiable than household or personal data, especially on a spontaneous basis, because the distribution of their characteristics is much more skewed. Also, in many business surveys, the largest businesses are selected with certainty. In some countries, databases of business data are more accessible, thereby enabling matching. In addition, many academic researchers might also serve as consultants to business and even bona-fide access to business microdata by these researchers might be incompatible with such consultant roles (they cannot be brainwashed of knowledge acquired in the course of their research). Moreover, countries may have issues of economic competitiveness (and possibly even security) due to sharing identifiable business data with researchers in other countries.

68. From the point of view of researcher access, the main differences between household or personal data and business data are that the dissemination streams that provide greatest protection are most relevant to business data.

69. In terms of the dissemination streams:

- statistical tables remain relevant, although the higher level of identification risk means that more detailed tables will generally not be available in respect of businesses;
- anonymised microdata files may only be relevant for the smallest businesses. For some research, small businesses may be a group of particular interest for researchers. Even then there will need to be ‘distortion’ of some data (e.g. financial data) to avoid matching with other databases (e.g. taxation data). An alternative is to present the data in ranges. Thus, anonymised microdata files are likely to be of limited use;
- for similar reasons, RAFs may only be relevant for microdata files of the smallest businesses. At least, use of these facilities will enable NSOs to control the matching risk, so it may not be necessary to ‘distort’ the data to protect confidentiality. But, if large businesses are included, it may be difficult to confidentialise outputs even if the researchers cannot directly access the microdata;
- data laboratory arrangements are likely to be most pertinent for access to microdata files of businesses. Such arrangements exist in Statistics Netherlands, for example (see Case Study 13).

70. Some research studies may be able to be supported with the consent of the businesses involved.

## VI. MANAGING TENSIONS BETWEEN NATIONAL STATISTICAL OFFICES AND RESEARCHERS

### How might the tension between NSO and researcher perspectives be resolved?

71. This will most effectively be done by NSOs moving from a risk avoidance strategy to a risk management strategy. How to do this is discussed in more detail in the following paragraphs.

72. There are definitely risks that have to be managed. The rapid expansion of databases, containing data about identifiable persons, means that it is virtually impossible to completely avoid identification through data matching of a significant number of persons even though names and addresses are removed, particularly if household structure is contained in the files. Many of these databases are held by the private sector where controls on their use are generally less stringent than for the public sector. Furthermore, technology advances have made data matching easier, whether by exact matching or statistical matching techniques (which can lead to exact matches in unique cases). Risk avoidance in essence means not allowing identifiable microdata to leave the premises of the NSO unless other steps such as data swapping or data perturbation are undertaken. (Note that risks will vary according to the size of country among other things. In smaller countries, the risk will be relatively higher because there are relatively more unique cases.)

73. Nevertheless, the microdata access provided by NSOs does not seem to have been an area of public controversy. Implicitly, there seems to be a reasonably high level of public acceptance of current practices although we are not aware of countries where there has been an extensive public debate. But general community concerns about privacy suggest there is a limit to what the public is likely to accept. A debate could be easily triggered (across national boundaries) by one unfortunate incident. The level of cooperation in statistical collections could be subsequently affected.

74. Transparency is important in order to avoid accusations of secrecy. Therefore, it is good practice for NSOs to be transparent in outlining that one of the valued uses of the data from some statistical collections will be to provide researcher access to confidentialised microdata under controlled conditions for specific purposes. This has to be managed carefully or the privacy advocates could sway public opinion. Support from respected and authoritative persons is very important.

### How do NSOs manage the risks of microdata access?

75. Some suggestions are outlined below:

- (i) agree on a set of principles that should be followed in the provision of access to microdata (such as those outlined in Chapter 3);
- (ii) ensure there is a sound legal and ethical base (as well as the technical and methodological tools) for protecting confidentiality. This legal and ethical base requires a balanced assessment between the public good of confidentiality protection on the one hand, and the public benefits of research on the other. A decision on whether or not to provide access might depend on the merits of specific research proposals and the credibility of the researcher, and there should be some allowance for this in the legal arrangements. Access should not be regarded as automatic;

- (iii) have an arms-length process for balancing these two public goods. It is good practice to set up an internal committee to debate these matters and make recommendations to the head of the NSO. Judgements are involved and ethics committees or similar bodies may be able to assist in situations where there is discretion in deciding whether to provide access or not. The public good arguments are much stronger if the results of the research are to be placed in the public domain;
- (iv) be completely transparent about specific uses of microdata to avoid suspicions of misuse;
- (v) be prepared to provide more access through remote access facilities and data laboratories if completely unidentifiable microdata for public release may not be possible without considerable distortion of the data. Explore other opportunities to use technological developments to improve access to microdata in such a way that adequate confidentiality protection is provided;
- (vi) pass some of the onus of responsibility to the research community. Ensure researchers understand the reasons NSOs are so protective of confidentiality. Ensure researchers are aware of the consequences to them and their institution if there are breaches. Follow up with appropriate retribution if there are breaches.

76. The last point requires some comment. The culture and value system of the research community is very different to that of an NSO. Researchers often regard some of the ‘controls’ inherent in the microdata access arrangements as unnecessary bureaucracy. Whilst there are no known incidents of researchers using their access to microdata to deliberately identify individuals, there have been incidents where microdata provided to them on an exclusive basis has been provided to other researchers without permission, or cases where microdata have been statistically matched without permission with other data to produce richer data sets. The researchers in question may feel they have done nothing wrong, as they have not tried to identify individuals. However, incidents of this type, if they become public, can undermine public confidence and should be treated seriously. NSOs and researchers operate in different cultures and take different views of risks from incidents. This has to be taken into account in the determination of procedures for release of microdata.

77. How can NSOs pass some of the risk back to researchers? Actions might include:

- (i) asking them to prove their bona fides as researchers and to demonstrate the public benefits of their research and that the microdata are necessary for this research;
- (ii) making them sign a legally binding undertaking with similar penalties to those operating for NSO staff if they breach confidentiality provisions;
- (iii) explaining the reasons NSOs are cautious. Ensuring researchers are fully aware of their obligations through appropriate education. Follow up with effective audit and monitoring procedures. It may be useful to establish a Code of Conduct in collaboration with the research community;
- (iv) where offences occur, withdrawing all current and future services from the researcher and possibly their institution for a period of time (e.g. until the institution has undertaken appropriate disciplinary action against the offender). Make them realise that the future release of microdata to any researcher may be at risk if there is strong public criticism. Undertaking legal action where appropriate.

78. The potential harm from an unauthorised disclosure cannot be underestimated, particularly if done deliberately. Such situations must be treated very seriously.

79. The reality is that a combination of legal, administrative and technical measures will be necessary to ensure public confidence in the arrangements. Furthermore, the research community must accept that it has no automatic right of access. The NSOs may be enabled to provide access but researcher access should be at the discretion of the NSO. There will be responsibilities associated with access. In particular, researchers should accept that they will have a shared responsibility to maintain and uphold the conditions under which they have been provided access. The limitations and safeguards may be more restrictive than exist with other data sets to which they have access but there is a good reason.

## **Other issues**

### Consent

80. It is sometimes argued that respondent consent should be sought before release of microdata outside of the NSO. They argue that respondents have a right to decide how their data should be used even if it is not identifiable. This should be discouraged, as:

- (i) there are significant practical issues associated with seeking and managing consent;
- (ii) data being provided are unidentifiable and are only being used for statistical purposes, consistent with the purpose of the data collection;
- (iii) it is very difficult to provide all the information required for a respondent to make a really informed decision, – and so many respondents will say “No” just as a precaution. The sample will soon become unrepresentative if it is reliant on just those who give consent.

However, there is an obligation, as stated elsewhere in these guidelines, to be transparent about the arrangements to respondents. By this means, it can be argued that passive consent has been obtained.

(Note: If allowed by law, informed consent would be appropriate in a situation where the publication of small aggregates allows users to infer the situation of a single sample unit (e.g. person or business) that is part of this aggregate. This situation is more likely to apply to business statistics.)

### Administrative data

81. There is another perspective on the consent issue. The data of a NSO can comprise data collected directly by themselves and data collected by administering authorities and passed on to the NSO. Unless there is specific provision in legislation or a protocol to the contrary, a NSO should not release data from administrative sources in microdata form without the consent of the administering authority (who may feel unable to give consent because of promises made to their respondents). Even when administrative data are already in the public domain, it would be courteous to advise the administering authorities to give them an opportunity to comment. Otherwise difficulties may arise with the supply of the administrative data. Administering agencies also have to manage their own privacy and confidentiality issues.

### Contingency Planning

82. It is important that NSOs do some contingency planning in the event the microdata access becomes an issue for public debate. They should not assume such a debate will not happen. What are some of the key defences?

- (i) NSOs can point to the care they take in providing confidentiality protection through devices such as anonymising the microdata, providing strong physical security protection and the care taken in devising a process for the assessment of the balance between the conflicting public goods of confidentiality protection and the public benefits of research.
- (ii) If an offence has occurred and a NSO is questioned, it should be open about the offences and the penalties that have been invoked; it should make clear that the breach is the responsibility of the researcher but that the NSO will take appropriate action in response to the breach.
- (iii) NSOs should point to the overall public benefits of providing microdata access, particularly for the situation where the offence has occurred, and give some good examples
- (iv) Well-known and respected people who are prepared to publicly support the arrangements should have been arranged. Senior privacy officials may be of particular importance in this regard.

## VII. MANAGEMENT ISSUES ASSOCIATED WITH THE RELEASE OF MICRODATA

### Managing decision making on confidentiality

83. There is always some chance of identification, even if very small. Software now exists which can estimate the proportion of records which are unique and therefore at some risk of identification.

84. It is the Chief Statistician or his or her delegate who needs to make the decision on the release of a microdata file, whether it be by an anonymised microdata file (public-use or licensed), through a remote access facility, or through a data laboratory. In order to make that decision, the Chief Statistician needs advice on whether, for example:

- the risk of identification is sufficiently small;
- the adjustments made to the data items have not unduly damaged the microdata file for research purposes; and
- the variables that have been collapsed are the most appropriate, taking into account both the needs of researchers and the identification risk.

85. As an illustration of the last point, choices can be made between the amount of detail provided on identifying variables such as geography, age and household structure.

86. Appropriate arrangements should be put in place to provide this advice on a consistent basis. It often needs to be supported by a research capability and could be located in a methodology area. Case studies 17 and 18 describe the arrangements in Slovenia and Australia respectively.

### Managing metadata

87. If users are to make effective use of microdata, they must have access to the appropriate metadata. This would include:

- (i) a description of the survey including any information on quality;
- (ii) a list of the data items and the classifications used (sometimes referred to as a 'data dictionary'); and
- (iii) definitions of the data items.

Provision of (i) will help ensure that the microdata are not used if the data are not really fit for the intended purpose.

88. As microdata are provided electronically, the metadata must be provided in a way that is accessible. If possible, the metadata should be released in association with the microdata. Printed copy may still be an effective means although the NSO web site is becoming increasingly useful for these purposes.

**Managing breaches by the researcher**

89. Efforts should be made to reduce the likelihood of breaches as outlined in the previous chapter. Nevertheless, breaches may occur and procedures for dealing with the breaches should be determined.

90. Breaches must be treated seriously. If this is not done, public confidence in the arrangements will erode. Also, breaches will occur more often if they are not treated seriously.

91. There are a number of ways of dealing with breaches. For example, if a legal offence has occurred, legal action should be considered. This is expensive but is essential in more severe cases to demonstrate the importance the NSO places on confidentiality, and reduce the likelihood of future offences.

92. Also, the researcher should be prevented from further access to microdata. This should be the minimum step that is undertaken.

93. Consideration should also be given to stopping further release to the institution of the researcher, at least until:

- (i) the institution has taken appropriate steps in dealing with the offence committed by the researcher; and
- (ii) the NSO is confident that the appropriate arrangements are in place within the institution to minimise the chance of further breaches.

94. The research community will generally be supportive of taking strong action against the relatively small number of offenders who may give the research community a bad name. It is in their long-term interest.

95. For minor breaches, a warning may be sufficient.

## VIII. SOME SPECIAL ISSUES

### International access

96. Cross-country comparisons are important for understanding the effectiveness of policies and programmes adopted by individual countries. Governments, in particular, find this useful for policy evaluation work. The benefits of providing access to anonymised microdata to researchers working on cross-country comparisons and international agencies are clear but there are also risks (e.g. passing on microdata without permission). With the exception of Eurostat, staff from international agencies are not subject to any national or international legislation other than the applicable staff rules of the organization. So some care has to be taken. The main difficulty is that there is more limited scope for retribution against breaches committed by staff of international agencies or researchers living in other countries. But, on the other hand, the probability of identification is much lower (as long as the researcher does not pass the microdata back to a third person in the host country). A further difficulty is that many countries do not have the legal authority to provide data to international agencies or researchers outside their country.

97. These guidelines have suggested moving towards a risk management rather than a risk avoidance approach in the provision of microdata. Risks are lower or perceived to be lower if the recipient agency has credibility and a trustworthy reputation. Risks are also more easily justified if the purpose for which the microdata are provided is in the interest of the country providing the data. This may be because the data are part of an international study being undertaken by an international agency or a respected international research collaboration (e.g. the Luxembourg Income Study).

98. Globalisation is increasing the importance of such international studies. It should be legitimate for NSOs to be able to support these types of studies through the provision of microdata. But it should be an enabling provision (i.e. the NSO should be able to decide whether to provide the data or not) that also specifies appropriate safeguards and conditions of release. NSOs should be more prepared to provide anonymised microdata where the risks are lower and the benefits are higher.

99. What are the options for researchers' access to data sets from other countries? How can international agencies obtain access to microdata for statistical and research purposes? The options include:

- (i) the data are collected directly by the international agency (or researcher) or through intermediaries (e.g. a specialist survey organization) in such a way that the forwarding of microdata is made known at the time of data collection;
- (ii) Public Use Files where they exist;
- (iii) Licensed Anonymised Microdata Files, where countries are able to do this;
- (iv) Remote Access Facilities with appropriate safeguards;
- (v) collaboration with a researcher based in the NSO or the NSO's country, who has access to the microdata.

100. From the point of view of microdata access, surveys of type (i) are preferable for international researchers. The PISA study is a good example. However, generally speaking, these data will not be collected under the statistics legislation operating in the various countries. The quality of the microdata, and particularly response rates, could suffer for some studies. This will depend on the type of study and the reputation of the organization undertaking the study. It also depends on the commitment of the data collectors in individual countries. This is a factor that international researchers need to consider before deciding to take this approach. There may be trade-offs between access and quality.

101. In some cases, it may be possible to satisfy the requirement of the study by providing very detailed data for analysis, but not microdata. This approach is used in the International Comparison Programme and the OECD/Eurostat Purchasing Power Parity study.

102. With respect to approach (ii), Public Use Files are only available for some countries. Licensed anonymised microdata files (i.e. approach (iii)) may be an option if not restricted by the legislation of the NSO. If NSOs are able to provide microdata in this way, factors that might be taken into account are:

- trust in the researcher and his or her institution;
- whether the study is of importance to the country; and
- whether such release would be consistent with promises made to respondents at the time of data collection.

103. It is likely that many countries will feel more comfortable releasing to particular international agencies or specific research projects rather than to the international research community more generally. Also, there may be some conditions that apply to particular researchers. For example, some countries may only feel comfortable releasing to researchers via the NSO of the researcher's home country. But in all cases of microdata release, it would be good practice to release only subject to specific conditions using either an undertaking or memorandum of understanding. Some of the conditions might include:

- limiting access to particular divisions of the international agencies and prohibiting the passing on of data to others;
- limiting the purposes for which the microdata could be used without permission;
- agreeing that the microdata must be returned on request (e.g. it may contain errors);
- allowing the opportunity to comment on published material which uses the microdata; and
- making clear the consequences of not following the conditions of release.

104. The most effective way of dealing with breaches of conditions of release will be to suspend further releases. The matter could also be taken up with more senior staff in the institution. For international agencies, this could be through diplomatic officials for more serious offences. But the key point is that breaches cannot be allowed to take place without some reaction. Otherwise, they will simply be repeated.

105. For many countries, the use of Remote Access Facilities (i.e. approach (iv)) may be the preferred route to provide access to international researchers. Under such arrangements, there are more controls, and the position of NSOs on international access to microdata is more easily defended if challenged. However, the usability of these arrangements for international access still needs to be improved. Experimentation is important.

106. There is another possibility. International researchers, including the international agencies, could work through networks of national researchers to achieve their aims (i.e. approach (v)) such as specified analysis against the microdata. Indeed, these national researchers could be located in the NSO for international studies of particular importance.

107. NSOs will need to decide whether they can provide access to international researchers or not, taking into consideration the range of issues discussed in this section. They should keep in mind that a risk management approach is being encouraged. For some research applications, the benefits may justify the risks involved, as long as the arrangement is legal. The risks may be lower for some institutions than for others. NSOs will also need to decide the most appropriate form of access in collaboration with the international researcher. To help ensure consistent decision making, countries should develop guidelines for access to international researchers and international agencies that are consistent with their own legislation. Or they may decide to change their legislation to enable access for justified cases.

108. Case Study 19 describes the arrangements used in the OECD PISA study and Case Study 20 describes a policy on international release of microdata for Australia.

### **Data linking**

109. The linking of data sets, whether by exact or statistical matching, can add considerable value to them. It can facilitate a much greater range of analyses. Health research, in particular, is an area where linked data sets can be of particular value. It is an appropriate function for NSOs to be involved in the linking of data sets for statistical purposes.

110. Increasingly, researchers are looking to utilise linked data sets that include links with the data sets of the national statistical office or other statistical agencies (including the population census in some countries). The statistical agency has to be the custodian for these linked data sets. There may also be situations where it is the preferred custodian of linked data sets even when the data sets come from outside the statistical agencies, because of the safeguards and public trust that already exist.

111. While there are clear benefits in data linking, there are also risks, particularly if the custodian of the linked file does not have confidentiality protection that often exists with the NSO. Identification risks are increased with linked data sets. Perceptions are also important. Studies in many countries show much public concern about linking databases. It is particularly important that the four principles outlined in Chapter 3 are followed for linked data sets.

112. In those countries where they exist, Privacy Commissions or equivalent bodies should support the arrangements for data linking.

113. Case Studies 21 and 22 describe how Canada and Sweden respectively manage data-linking arrangements in their countries.



## ANNEX 1.1. CASE STUDY

### LEGISLATION TO SUPPORT RELEASE OF MICRODATA - AUSTRALIA

#### 1. Broad description

This is delegated legislation (referred to as a ‘Ministerial Determination’, but in effect a regulation) that enables the Australian Bureau of Statistics (ABS) to release microdata to approved users for statistical purposes. It also outlines the conditions of release and the penalties for any breach of those conditions.

#### 2. Why is it good practice?

It provides a degree of certainty to both the ABS and the potential users of microdata about the arrangements for release. The legislation also outlines the arrangements that the Parliament is happy with. As they are enshrined in delegation legislation, they are in the public domain.

#### 3. Target audience

Primarily the research community who are the main users of microdata.

#### 4. Detailed description

The specific legislation is outlined in Part 5. There is also a supporting statement providing policy, rules and guidelines to assist ABS staff involved in the release of microdata.

Each new release of microdata requires the approval of the Australian Statistician in view of the potential sensitivity of releases. Each release to individual clients requires the approval of a senior manager, employing the delegated authority of the Australian Statistician.

A Microdata Review Panel has been established to provide advice to the Australian Statistician on microdata releases, particularly the steps that need to be taken to ensure the release complies with the confidentiality test imposed by the legislation.

#### 5. Supporting legislation

Disclosure of unidentified information:

(1) Information in the form of individual statistical records may, with the approval in writing of the Statistician, be disclosed if:

- (a) all identifying information such as name and address has been removed;
- (b) the information is disclosed in a manner that is not likely to enable the identification of the particular person or organization to which it relates; and
- (c) the Statistician has been given a relevant undertaking by each person required by sub-clause (2) to give a relevant undertaking in relation to the information.

(2) The persons required to give a relevant undertaking are:

- (a) for information to be disclosed to an individual, the same individual; and
- (b) for information to be disclosed to an official body:
  - (i) the responsible Minister in relation to, or a responsible officer of, the official body; and

- (ii) if the Statistician considers it necessary in a particular case, each individual in the official body who will have access to the information.
  - (c) for information to be disclosed to an organization other than an official body:
    - (i) a responsible officer of the organization; and
    - (ii) if the Statistician considers it necessary in a particular case, each individual in the organization who will have access to the information.
- (3) In this clause: ‘relevant undertaking’ means an undertaking in writing that use of the information in relation to which the undertaking is given is subject to the following conditions:
- (a) no attempt will be made to identify particular persons or organizations to which the information relates;
  - (b) the information will be used only for statistical purposes;
  - (c) for information to be disclosed to an individual, the information will not be disclosed to anyone without the approval in writing of the Statistician;
  - (d) for information to be disclosed to an official body or other organization:
    - (i) the information will not be disclosed to anyone outside the body or organization without the approval in writing of the Statistician; and
    - (ii) if the Statistician considers it necessary in a particular case, the information will not be disclosed to an individual in the body or organization who has not given a relevant undertaking;
  - (e) if the Statistician considers it necessary in a particular case, either or both of the following:
    - (i) the information, and all copies (if any) of the information, will be returned to the Statistician as soon as the statistical purposes for which it was disclosed have been achieved;
    - (ii) access by officers to information, documents or premises will be given as may be necessary for the purpose of conducting a compliance audit concerning observance of the conditions under which the information is disclosed;
  - (f) any other condition that, in the opinion of the Statistician, is reasonably necessary in a particular case.

In a different part of statistics legislation, it is made clear that a person who fails to comply with an undertaking, as prescribed in (2) above, is guilty of an indictable offence punishable on conviction by a fine not exceeding \$5000 or imprisonment for a period not exceeding 2 years, or both.

## 6. Strengths

- (i) Provides a basis for arrangements that are understandable by both the NSO and researchers.
- (ii) Provides for significant penalties for legal breaches; this may be a reason why no known breaches have occurred.

- (iii) Microdata protection is partly provided by a legally enforceable undertaking. This means that some protection (e.g. prevention of matching) can be provided through the undertaking.
- (iv) Provides wider access to the data than would otherwise be the case, thereby achieving a greater return on the high investment in data collection and respondent burden.
- (v) Provides statutory authority and transparency for release practices and a basis for the public defence of those practices.

## **7. Weaknesses**

- (i) Researchers still believe the conditions of release are too limiting i.e. the steps taken to make the data confidential result in too much of the detail not being released.
- (ii) Disclosure of microdata, even under circumstances demanding strict confidentiality, can alarm the privacy constituency and in the worst case, have the potential to impact on response rates.
- (iii) Compliance with the limitations and conditions imposed by legislation can impose an administrative burden on both the NSO and the researchers, delay the release of information, restrict the range of researchers who can have access to the information, restrict the uses to which the information can be put and limit the nature of the information which can be released.

## **8. References**

The supporting statement referred to in Part 4 is available on request from [teresa.dickinson@abs.gov.au](mailto:teresa.dickinson@abs.gov.au)

## ANNEX 1.2. CASE STUDY

### LEGISLATION TO SUPPORT RELEASE OF MICRODATA - FINLAND

#### 1. Broad description

Legislation (the Statistics Act) enables Statistics Finland to release microdata to approved users for scientific research and statistical purposes. It also outlines the conditions on the release and the penalties for any breach of these conditions.

#### 2. Why is it good practice?

The law sets out conditions under which microdata can be released. Written guidelines by Statistics Finland give further directions and assure equal treatment of all applicants. The law gives a certain leeway to Statistics Finland in assessing the threat to confidentiality that the data might impose. The principles of data release are well known by all parties (statistical authorities, data suppliers and data users). The practice does not weaken data suppliers' trust in the confidentiality of basic data. The decisions on access to microdata are made solely by the statistical authority.

#### 3. Target audience

A licence to use data may be issued to an official body, an institution or a person in charge of research. In cases where the licence is issued to an official body or an institution, it is granted to a specific person or specific persons.

#### 4. Detailed description

The essential principles and procedures of data release are prescribed in the Statistics Act. Statistics Finland has given more detailed guidelines on data release. These guidelines and the application form for access to microdata are publicly available on Statistics Finland's website.

An application for a licence to use data must be submitted in writing. The applicant must specify the purpose for which the statistical data are to be used, the material requested from Statistics Finland and any other data that will be used. A research plan should be appended to the application.

When considering the granting of a licence to use basic data, Statistics Finland first determines whether the data can be processed in-house to obtain the statistics requested by the applicant. In considering the application, account is also taken of the possibility that the applicant will obtain reliable results on the basis of this material. Particular attention is also paid to data protection issues.

Statistics Finland also takes into consideration any other data that the applicant may already have at their disposal. If a licence is granted and the data in the possession of Statistics Finland are to be combined with other data, this combining must take place at Statistics Finland, which shall remove all identification variables from the combined material.

Statistics Finland will not grant a licence for data covering the whole country or a whole region. Authorisation to use entire data files is generally granted only in exceptional cases, such as specific research purposes and where the material does not contain sensitive data.

Before releasing the data, all identification information is removed from the material for which a licence has been granted, or the data are made less detailed or combined with other data in order to prevent identification. The information to be removed or the manner of combination shall be indicated in the licence decision. Information on age, gender, education and occupation may be released with identification data if the applicant is entitled to collect such data by virtue of the Personal Data Act. This must be indicated in the application. An additional requirement is that the release of these data in identifiable form is considered essential to the study.

The decision to grant access to microdata is made by the Director General when data are to be released abroad. In other cases the decision is made by the director of a statistics department. The licence is granted for a limited period only. Statistics Finland has a Committee on Statistical Ethics which helps decision-makers by giving opinions on complicated data release issues and on all cases where data are to be released abroad.

Each licence is accompanied by the terms and conditions applicable to the use of the data. The data may only be used for the purpose indicated in the decision. The data shall be treated as confidential and may not be handed over to others without authorisation from Statistics Finland. No attempt must be made to identify the data subjects from the material and the data must be destroyed by the set date.

## **5. Supporting legislation**

All statistical data are confidential irrespective of the data source. The release of confidential data is determined by the Statistics Act (Section 13).

The data obtained by a statistical authority for statistical purposes may only be released to a third party on terms laid down in the Statistics Act or in another act concerning especially the National Statistical Service or upon express consent of the subjects of the data.

Confidential data collected for statistical purposes may be released for use in scientific research or statistical surveys concerning social conditions. Such data may not be released for use in an investigation, surveillance, legal proceedings, administrative decision-making or other similar handling of a matter concerning an individual, enterprise, corporation or foundation.

Identification data may not be released. Both direct and indirect identification of personal data must be prevented. As far as other data (e.g. business data) are concerned it is sufficient to prevent direct identification. However, access to business microdata is usually granted only at the premises of Statistics Finland.

The decision to grant access to microdata is always made by the statistical authority concerned. When making the decision, data protection issues must be taken into consideration.

Violation of statistical confidentiality is a punishable offence (Section 24 of the Statistics Act). The punishment may be a fine or imprisonment not exceeding two years.

## 6. Strengths

Legislation provides a clear basis for arrangements that the National Statistical Institute, data suppliers and researchers can understand.

Data obtained from administrative sources can be released without the permission of the data supplier. This makes the procedures of data release simpler and enables the use of very large data files.

Legislation ensures that the release of data collected for statistical purposes cannot be regulated by any other act than the Statistics Act. This enhances the trust of data suppliers.

Legislation prescribes severe punishments for breaches of law.

## 7. Weaknesses

From Statistics Finland's perspective:

- Making data non-identifiable demands a great amount of work, which increases their cost to researchers.

From data users' perspective:

- Data are often regarded as expensive by researchers.
- Researchers sometimes think that there are too many restrictions to obtaining and using data.

## 8. References

The legislation can be found at [http://tilastokeskus.fi/org/index\\_en.html](http://tilastokeskus.fi/org/index_en.html)

## ANNEX 1.3. CASE STUDY DATA CUBES - NETHERLANDS

### 1. Broad description

Statistics Netherlands (*Centraal Bureau voor de Statistiek*, or CBS) releases its publishable information in its output database StatLine on the Internet, [www.cbs.nl/en/statline](http://www.cbs.nl/en/statline). Usually, this information takes the form of multidimensional tables, or data cubes. These tabulations are safe from the perspective of statistical confidentiality protection. The user selects and processes his own views on these data cubes. Occasionally, statistical work is commissioned and paid for by third parties, resulting in data cubes.

### 2. Why is it good practice?

Data cubes are the main vehicle for releasing all statistical information. Statistical confidentiality protection is applied in a routine fashion. Moreover, data cubes can be easily linked and compared on a meso level. Conversely, a lack of coherence is easily discovered. Adding data cubes to the StatLine database ensures that statistical information is produced and published to serve the public at large.

### 3. Target audience

Data cubes are primarily made and used to serve the public at large. Even if they are produced and paid for by a third party, as a matter of policy the resulting data cubes are available for all.

### 4. Detailed description

Statistics Netherlands has published several papers on the art of ‘cubism’.

### 5. Supporting legislation

Three sections of the Statistics Netherlands Act ([www.cbs.nl/en-GB/menu/organisatie/statistics-netherlands-act.htm](http://www.cbs.nl/en-GB/menu/organisatie/statistics-netherlands-act.htm)) are relevant, pertaining to its general public task, its commissioned work, and the precondition of statistical confidentiality.

Section 3 states that it is the legal task of Statistics Netherlands “to carry out statistical research for the government for practice, policy and research purposes and to publish the statistics compiled on the basis of such research”.

According to section 5, “CBS may occasionally carry out statistical work for third parties.”

Section 37 reads:

“1. The data received by the director general in connection with the performance of his duties to implement this act shall be used solely for statistical purposes.

2. The data referred to in the first subsection shall not be provided to any persons other than those charged with carrying out the duties of the CBS.

3. The data referred to in the first subsection shall only be published in such a way that no recognisable data can be derived from them about an individual person, household, company or institution, unless, in the case of data relating to a company or institution, there are good reasons to assume that the company or institution concerned will not have any objections to the publication.”

## **6. Strengths**

Statistics Netherlands is in full control as far as statistical disclosure protection is concerned. The user can rely on the professional quality of the statistical information. It is rewarding for staff to produce information that is in demand. The data cubes make it ever easier to relate various bits of statistical information to each other. Experiences with commissioned work may be fed back into the standard statistical programme as an indication of user preferences.

## **7. Weaknesses**

By definition data cubes are less informative and less flexible than microdata (unit level records) for researchers. As commissioned work has to be paid for, data cubes may appear to be expensive for the commissioning party.

## ANNEX 1.4. CASE STUDY

### PUBLIC USE MICRODATA - UNITED STATES

#### 1. Broad description

The U.S. Census Bureau first published public-use microdata for the 1970 Decennial Census. Microdata files of decennial censuses have been released since then, as well as public use microdata files from selected demographic surveys. The Census Bureau does not produce public use microdata from its economic censuses and surveys.

In the mid 1980s the Census Bureau established a Microdata Review Panel to oversee the content of microdata publication. This included ensuring that microdata files met disclosure avoidance conditions. In the mid 1990s, the Microdata Review Panel was replaced by the Disclosure Review Board (DRB), with a greater emphasis on disclosure avoidance. By this time, microdata were the primary publication form for servicing the Census Bureau's more sophisticated public users. Because Census Bureau data products that are released to the public are available to all users, the role of the DRB is to establish disclosure avoidance guidelines for all of the Census Bureau's data products (including microdata) and to ensure that they adequately protect the identity of individual respondents. In practice, a checklist approach is used to assess these data sets. In addition, ongoing research is conducted to ensure that disclosure avoidance techniques are consistent with current conditions.

#### 2. Why is it good practice?

Microdata publication changes the role of the NSO, largely eliminating any interpretive function. The NSO is able to accommodate more interests and maintain itself as a neutral party. Interpretation of the data becomes more robust as more parties are able to examine the data in detail.

#### 3. Target audience

All users, from sophisticated analysts for micro-simulation modelling and policy evaluation to federal, state, and local governments, academic researchers, market researchers, private businesses, and the general public.

#### 4. Detailed description

The Census Bureau has published microdata files from decennial censuses since 1970. The medium of publication for the 1970 and 1980 Public Use Microdata Samples (PUMS) was mainframe tape. The 1990 Census Public Use Files are available on both tape and CD-ROM. Census 2000 microdata are available via CD-ROM and the Internet. Changes in media and technological advances have led to broader access by users in general and by type of user in particular.

For Census 2000 two principal sets of public use files were released – the 5-Percent PUMS and the 1-Percent PUMS. The two sets are mutually exclusive. The 5-Percent file contains data for 5% of all households in the country, is released for public use microdata areas (PUMAs) of at least 100,000, and requires the PUMAs to follow state boundaries. The 1-Percent file contains more detailed characteristics data for 1% of all households and is based on superPUMAs of at least 400,000 that do not cross state boundaries.

In addition to decennial census information, the Census Bureau public-use microdata products, provided through the Internet (FTP) and CD-ROM, include the following ongoing surveys:

- Current Population Survey (CPS);
- Survey of Income and Program Participation (SIPP);
- American Housing Survey (AHS);
- Survey of Program Dynamics (SPD);
- American Community Survey (ACS); and
- Consumer Expenditure Survey.

Personal identifiers are removed from these files and only large geographic areas are identified on microdata records. The Census Bureau uses a basic population threshold of 100,000 in conjunction with other methodologies, to avoid disclosure. Many of the surveys for which Public Use Files are produced use a larger geographic unit (in terms of population) in order to offer more detailed data. To further protect confidentiality, there is limited detail on items such as place of residence, place of work, high incomes, and others. (See Zayatz (2002), for more detail about disclosure avoidance methods used for the Census 2000 PUMS.)

## 5. Supporting legislation

The Census Bureau's authorizing legislation is Title 13, United States Code. Section 9(a)(2) of this law prohibits the Census Bureau from making "any publication whereby the data furnished by any particular establishment or individual under this title can be identified." At the same time, the law states that the Census Bureau is encouraged to make "statistical use" of the data in its possession. Although some thought has been given to offering licensed access to microdata, as a means of expanding access to advanced users while ensuring enhanced protection of the data, legal interpretation of the Census Bureau's statute suggests that this is not an option. According to the Census Bureau's legislation, the data either are public or they are not – if they are public then they must be made available to any user; if they are not, they may only be accessed by persons who have taken the Census Bureau's Oath of Nondisclosure, who use the data only for statistical purposes, and are subject to severe penalties for disclosure.

In the United States, each agency has its own legislation and many statistical agencies do not have specific confidentiality protection as part of their statute. In 2002, the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) was passed, which guarantees that data collected under the CIPSEA with a pledge of confidentiality must be kept confidential, subject to severe penalties for disclosure, and which ensures that data collected for statistical uses may not be used for administrative or compliance purposes. This new legislation helps protect microdata that may be released by other U.S. federal agencies.

## 6. Strengths

For many data users, the summary tables and tabular and narrative profile reports released meet their needs. Microdata are released for advanced users who want to create or define their own tabulations, to be able to further draw on the richness of detail recorded in the census or survey.

Census Bureau microdata files are available to the general public without restriction on their use, and while the Census Bureau offers limited access to non-public microdata for selected users at its Research Data Centers, the ability to obtain public use microdata files permits users to access these rich data sets in their own settings, without the need for Census Bureau oversight.

## 7. Weaknesses

The methods used to make the data disclosure-proof can be damaging to some characteristics of interest:

- Geography is largely suppressed;
- Variables pertaining to collection are seldom included; and
- Data are being suppressed more often due to the presence of overlapping external data. This problem is likely to worsen.

Unfortunately, the more sophisticated the disclosure avoidance techniques are, the less undisturbed data can be released, ultimately affecting analysis, often in unknown ways. Recent advances in computer technology and data mining techniques increase concerns about the ability to continue to release detailed microdata files, and better methods are needed to measure microdata disclosure risk and the bias added by disclosure avoidance techniques.

## 8. References

Doyle, P. et al. (eds) (2001) *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*; North-Holland.

Duncan, George T; Jabine, Thomas B.; and de Wolf, Virginia A (eds.) (1993) *Private Lives and Public Policy*, Committee on National Statistics Panel on Confidentiality and Data Access, National Academy Press, Washington, DC.

Federal Committee on Statistical Methodology (1994) *Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology*, Office of Management and Budget: Washington, DC.

U.S. Census Bureau (2003) *Access to Microdata – Issues, Organization and Approaches*, Conference of European Statisticians, Geneva, June 10-12, 2003.

Zayatz, L. (2002) “SDC in the 2000 U.S. Decennial Census”, in *Inference Control in Statistical Databases* (Josep Domingo-Ferrer, ed), Springer.

## ANNEX 1.5. CASE STUDY RELEASE OF ANONYMISED MICRODATA FILES (LICENSED FILES) - AUSTRALIA

### 1. Broad description

Anonymised microdata files (licensed files) in Australia are referred to as Confidentialised Unit Record Files (CURFs). Key measures undertaken by the Australian Bureau of Statistics (ABS) to protect the data are: requiring anyone who uses the data, and the organizations that employ them, to sign an undertaking with the ABS; obtaining user commitment to confidentiality principles; and perturbing data or reducing detail on files to make it very difficult for units to be identified. CURFs are most commonly made available to users either on a CD-ROM or through a remote access data laboratory (RADL™). CURFs available on RADL™ contain more detail than those on CD-ROM. In selected cases users may have access to a CURF through an on-site data laboratory.

### 2. Why is it good practice?

Releasing microdata in this manner constitutes good practice as:

- perturbation of data and masking of records is undertaken to maintain the integrity of the data while protecting the confidentiality of an individual's data; and
- placing restrictions on how the data are used, as set out in a legal undertaking to be signed by each user and their organization, ensures both the user and the organization accept responsibility for keeping the data confidential and secure.

### 3. Target audience

CURFs are aimed at Australian researchers and analysts within government, academia and other non-government organizations, who seek to undertake more in-depth analysis than is possible using tabular aggregated data.

CURFs are not generally released to overseas applicants. In very selected instances the ABS allows overseas researchers to access CURFs via the RADL™ if they are sponsored by a suitable Australian organization. The sponsoring organization is required to sign an undertaking with the ABS.

### 4. Detailed description

The ABS has adopted a manner of release for CURFs that protects the data in three ways: confidentialising the unit record file to control the detail available; providing modes of access appropriate to the level of detail available; and requiring users and organizations with access to the data to sign an undertaking that restricts how they use the data.

The unit record files are confidentialised by removing name and address information, by controlling and limiting the amount of detail available, and by perturbing or deleting data where it is likely to enable identification of individuals.

Each CURF release is personally approved by the Australian Statistician, following advice from a Microdata Review Panel consisting of three senior executives. The panel makes a detailed assessment of each CURF to ensure that the disclosure risk is low.

There is protection inherent in the different access modes and in the different levels of data provided in each. The ABS provides three different modes of access for CURFs - CD-ROM, the RADL™ and the ABS Data Laboratory (ABS DL). CURFs available on CD-ROM are labelled Basic CURFs and are restricted to a relatively small number of variables released in broad categories. RADL™ users can also access Expanded CURFs that contain more variables and more detail, with extra protection provided by the automatic logging of RADL activity and subsequent audits of this activity. Specialist CURFs contain the most variables and detail and can only be accessed via on-site ABS Data Laboratories.

Each user must apply to be granted access to a CURF, explaining their intended use of the CURF. Both the User and a Responsible Officer of the employing organization must sign a legal undertaking in which they agree:

- access to information about individuals will be restricted to officers of the organization who have signed an individual undertaking with the ABS;
- users will not attempt to identify individuals;
- users will not match the unit data to other files of unit data;
- ABS officers are allowed access as necessary to audit compliance with these rules;
- CURF usage is limited to the specified and approved individual ‘Statistical Purpose’; and
- any sensitive printed data and output will be stored in a secure place.

The organization must monitor its officers that have access to the CURF and ensure that all have signed an individual undertaking with the ABS. Access to CURFs are for statistical purposes within an organization. If an individual changes organizations, they must surrender access and notify the ABS.

The responsible officer is generally the head or deputy head of an organization, department or university. They are required to sign an undertaking about the storage and use of the CURF. Breaches can be addressed by sanctions against both the individual user and the organization as well, including removal of access to all microdata for all individuals in the organization.

## 5. Supporting legislation

The release of microdata by the ABS is governed by legislation; namely, the Census and Statistics Act 1905. This legislation enables the Australian Statistician to release unit record data, provided this is done “in a manner that is not likely to enable the identification of a particular person or organization to which it relates.” Details are provided in Section 5 of Annex 1.1.

## 6. Strengths

- (i) Allows for a range of access mechanisms to suit a range of uses.
- (ii) Allows for access to more detailed data to be granted to users who are able to work with a greater level of environmental protections.
- (iii) Microdata protection is partly provided by a legally enforceable undertaking. This means that some protection (e.g. prevention of matching) can be provided through the undertaking.
- (iv) Sanctions can be applied against users and organizations that breach the undertakings, providing additional motivation to ensure data access and use is appropriate.

## 7. Weaknesses

- (i) Researchers still believe the protections applied directly to the microdata are too limiting. They believe too much of the detail is not being released, especially for some of the most identifiable sections of the population (e.g. large households).
- (ii) It is more costly to support a range of access mechanisms than a single access mechanism.

## 8. References

The Census and Statistics Act 1905 <http://scaletext.law.gov.au/html/pasteact/1/580/top.htm>

The Statistics Determination 1983 <http://scaletext.law.gov.au/html/pastereg/0/414/top.htm>

CURF undertakings & the Responsible Access to ABS CURFs Training Manual is at: [www.abs.gov.au/websitedbs/D3110129.NSF/85255e31005a1918852558ac00697645/72d92417a0ba71b5ca256d01002c47a4!OpenDocument#Untitled%20Section\\_6](http://www.abs.gov.au/websitedbs/D3110129.NSF/85255e31005a1918852558ac00697645/72d92417a0ba71b5ca256d01002c47a4!OpenDocument#Untitled%20Section_6)

## ANNEX 1.6. CASE STUDY

### RELEASE OF LICENSED MICRODATA FILES - NETHERLANDS

#### 1. Broad description

For its social sample surveys Statistics Netherlands (*Centraal Bureau voor de Statistiek*, or CBS) releases about ten standard microdata files each year. The microdata are protected against disclosure but not to the last detail. The remaining risk is dealt with by a contract (or license). The microdata are available to legitimate researchers. They are released on tape or on disk, usually in the SPSS format.

#### 2. Why is it good practice?

The community of social researchers is quite large. By releasing standard microdata files from its social sample surveys CBS serves this community. A basic acquaintance with SPSS is widespread among them.

Research on these files:

- reduces expenditure of tax payers' money on data collection efforts;
- reduces response burden;
- provides researchers with readily available microdata;
- turns CBS files and corresponding definitions into a *de facto* standard;
- provides end users within the policy domain with high-quality information within a short turnaround time.

Social sample survey microdata are relatively easy to protect against disclosure.

Two of the national 'planning offices', or independent government research institutes (SCP and CPB) and some five university faculties have a full subscription to these licensed microdata files. In addition over 70 files are released to individual institutes and researchers.

#### 3. Target audience

Microdata are released under a contract or license to legitimate researchers only. Section 41 of the law mentions the researchers that are qualified. Amongst them are universities and other research institutes with a legal foundation, but also Eurostat and NSOs within the EU. A residual category of applicants must be formally admitted by the Central Commission for Statistics (CCS), the supervisory body for CBS. The CCS has set its own selection criteria and procedures, in which a focus on statistical (aggregate) research, independence from administrative authorities, and the intention to share results in the public domain are predominant. The CCS has no objection in principle against admitting non-EU universities, for example. A commercial bank or a journalist would not be eligible, however.

#### 4. Detailed description

Microdata files are released, nowadays usually on CD-ROM, to interested researchers that are qualified according to the law or to the CCS. The files are compiled and documented from the social sample surveys carried out by CBS. Of course, they do not contain formal identifiers or matching numbers. Other identifying variables are collapsed or protected in other ways. The sampling factor (1% for the Continuous Labour Force Survey being the maximum) by itself protects respondents.

#### 5. Supporting legislation

Providing microdata to researchers is legally defined as an exception to the general obligation of statistical confidentiality. The general obligation reads as section 37:

- “1. The data received by the director general in connection with the performance of his duties to implement this act shall be used solely for statistical purposes.
2. The data referred to in the first subsection shall not be provided to any persons other than those charged with carrying out the duties of the CBS.
3. The data referred to in the first subsection shall only be published in such a way that no recognisable data can be derived from them about an individual person, household, company or institution, unless, in the case of data relating to a company or institution, there are good reasons to assume that the company or institution concerned will not have any objections to the publication”.

The microdata release policy is supported by section 41:

- “1. Contrary to the provisions of Section 37 the director general may, on request, provide or grant access to a set of data to a department, organization or institution as referred to in the second subsection for the purposes of statistical or academic research where appropriate measures have been taken to prevent identification of individual persons, households, companies or institutions from those data.
2. A set of data as referred to in the first subsection may be provided to or made accessible to:
  - a. a university, within the meaning of the Higher Education and Research Act;
  - b. an organization or institution for academic research established by law;
  - c. planning offices established by or by virtue of the law;
  - d. the Community statistical agency and national statistical agencies of the member states of the European Union;
  - e. research departments of ministries and other departments, organizations and institutions, in so far as the CCS has given its consent.

The importance of statistical confidentiality is apparent in section 42:

“The director general shall only grant a request as referred to in Section 41 if the director general considers that the applicant has taken adequate measures to prevent the set of data being used for purposes other than statistical or academic research.”

The remaining risk of disclosure (considering the adequate measures mentioned in section 42) is dealt with by a contract or license. The contract is signed by the institute that has requested access to the microdata. An appendix to the contract is a confidentiality statement to be signed by each individual researcher with access to the data. There is no legal punishment or fine in the case of a transgression of legal or contractual obligations of confidentiality.

The research community itself has drafted, and agreed upon, codes of conduct for the social and epidemiological sciences. These codes have been accepted by the national privacy authority CBP. They may be interpreted as a sign of awareness on the side of researchers of ethical and legal problems of privacy and confidentiality. One of these codes installs a commission of appeal for respondents, on which a staff member of CBS serves.

Apart from supporting legislation there has been since 1994 a long-term contract with the Netherlands National Science Foundation (NWO). As a broker, it couples data providers, first and foremost CBS, and data users, primarily (but not exclusively) with a focus on the universities. Under this long-term contract, CBS obliges itself to make available at least eight social sample survey microdata files each year. NWO pays £450,000 per annum. Concrete users of microdata pay an additional small fee (varying from £1,000 to £5,000 depending on the size of the file, with a discount for older files, as well as for the full package for a whole year). NWO also organises publicity, user consultation days (on specific files or themes), and an independent formal evaluation every four years.

## 6. Strengths

The researcher can process and analyse the microdata at his own computer, at his own time, with his own familiar and specialised software.

From the statistician’s perspective an initial investment is needed for preparing and documenting the microdata file. But from then onwards, efforts can be quite minimal. The more microdata are used, the more value is added for society at large and for research in particular. Furthermore, the data quality and documentation is enhanced by feedback from intensive use.

## 7. Weaknesses

From the researcher’s perspective the microdata do not always contain sufficient detail. In some cases microdata are even deemed too sensitive to be allowed to leave the statistical office at all. For example, microdata from businesses, fiscal income statistics and causes of death statistics are not permitted to leave the office, which is an extreme example of collapsing the microdata.

Because of the lack of formal identifiers and the collapsing of some of the background variables, the microdata can not flexibly be expanded with new variables.

Some of the contractual conditions (CBS may want to screen draft publications or to inspect the ICT facilities on which the researchers have access to their data) are sometimes interpreted as ‘organised distrust’ if not outright ‘strangulation’ by CBS of research.

Some statistical staff consider it a threat that others use ‘their’ microdata and publish results that should have been on the official statistical programme.

## ANNEX 1.7. CASE STUDY

### RELEASE OF LICENSED MICRODATA FILES - SWEDEN

#### 1. Broad description

These are the arrangements for Statistics Sweden's release of confidential microdata (licensed microdata files) to approved users for statistical and research purposes.

#### 2. Why is it good practice?

The arrangements ensure that the release of confidential data are in accordance with the legislation concerning confidentiality and protection of individual's privacy. The legislation, decided by parliament, provides the limits for release of data, e.g. research purposes, and legally underpins and constitutes administrative and technical safeguards.

#### 3. Target audience

Statistics Sweden mainly provides access to microdata to public authorities and people or organizations performing scientific research (universities and research institutions). Statistics Sweden also provides access to microdata to other authorities and municipalities producing statistics.

#### 4. Detailed description

According to the main principle, confidential data may be released to a third party only for the purpose of statistics production, statistical analyses and research. The legislation is outlined in Part 5.

Statistics Sweden provides access to data which do not allow direct or indirect identification of individuals or of other data subjects like enterprises. This means in practice anonymous data or data without name, address and identification number. Both the anonymous and the de-identified data are in principle only available to the researcher for a specified period, for a specified project and for access by specified staff of an institution.

In addition to laws and regulations on data confidentiality, Statistics Sweden follows a screening procedure requiring a written application from the researcher. In the application the researcher is required to describe the project, variables and periods during which data are used in the research, and also specify the people taking part in the project. If the project involves processing of sensitive personal data the researcher is required to add the approval of a research committee.

Heads of Statistics Sweden's departments are the only ones who can decide on the release of confidential data. Furthermore, the Director General shall always decide on matters that are of fundamental importance. An advisory committee has been established to provide advice in difficult cases or matters of principle.

When microdata are released to a researcher at a private institute or organization, Statistics Sweden imposes legal restrictions limiting the researcher's right to pass on or use the information. If data are released to a researcher in another authority (e.g. a university), data will also be confidential at the authority receiving data, according to the Swedish Secrecy Act. In addition, researchers at other authorities normally sign a general confidentiality statement when receiving the data.

When microdata are released, it is common that a pseudo-identifier replaces the identification number. If the user needs annual series of microdata information for the same individuals, the pseudo-identifier may be connected with the identification number, and the combination is to be stored by Statistics Sweden. The possibility of having new information added by storing the combination of pseudo-identifier and identification number is restricted to research projects and for statistical purposes.

The main method of giving access to microdata for research has been to deliver the data to the user by sending a micro disc by post. The data and the metadata are sent separately.

However, since 2005 it has been possible for researchers to get access to microdata online (remote access). This allows researchers to have online access to specific servers at Statistics Sweden. However, all data processing will be carried out on the server at Statistics Sweden and no downloads are allowed. The results are frequently sent by e-mail to the researcher in the form of tables. The system is similar to the remote access system of Statistics Denmark. The Danish system is further described in the case study from Denmark.

Supply of microdata is not covered by grants provided in the state budget for production of statistics. Consequently, the costs involved in supplying microdata are to be paid by the customers. The principle is that full costs should be reflected in providing the data, i.e. cost recovery. The costs involved are to cover not only direct labour costs, but also overhead such as rental of the premises, office costs, staff costs, EDP costs, marketing costs, development costs and a proportion of the joint costs of the statistical institutes for management and administration.

## **5. Supporting legislation**

According to the Swedish Secrecy Act, any information concerning personal or economical circumstances of private subjects shall be confidential if it is managed within an authority responsible for producing statistics. However, information needed for statistics or research purposes and information which is not directly related to the private subject may be disclosed, if it is evident that the information can be disclosed without the person whom the information concerns suffering loss or other harm.

The obligation of confidentiality will – according to the law or by imposition of a duty of non-disclosure – also apply to the recipient of the data. Breach of confidentiality restrictions is punishable.

However, it is not possible to impose restrictions when data are released to other authorities. It is therefore also important for Statistics Sweden to take into consideration if data will be confidential according to the Swedish Secrecy Act at the authority receiving data.

The Statistics Act regulates the use of statistical information. According to this act, data collected for statistical purposes in accordance with any prescribed obligation to provide information, or which is given voluntarily, may in principle only be used for the production of statistics. From this principle there are exceptions that make it possible to provide access to data for research purposes and public planning. However, a condition for the use for research is that there should be no incompatibility between the purpose of such processing and the purpose for which the data was collected. The processing of data, which includes release of data, must also be in accordance with the regulation concerning the protection of the individual's privacy.

A scientific project involving processing of sensitive personal data without consent is subject to notification to, and approval by, a research committee before such processing can commence. This applies to all surveys, whether conducted by a public administration, individuals or enterprises. If the committee approves the processing, personal data may be released and used in research projects unless otherwise provided by the rules on confidentiality. This means that the statistical office may take other issues into consideration even if the research committee has approved the processing of data.

## **6. Strengths**

The arrangements are understandable by all parties, and transparent.

## **7. Weaknesses**

Researchers sometimes point out that the handling of their requests takes too much time and take the view that Statistics Sweden is too restrictive in releasing microdata.

The framework mainly depends on public confidence in research institutions and researchers. When microdata are released outside Statistics Sweden it is not possible for staff to control the use of the data. The Swedish Data Inspection Agency may observe illegal use of NSO data on their inspections at the institution.

## **8. References**

*The Official Statistics of Sweden – Annual Report 2004*. This report includes relevant Swedish legislation.

[www.scb.se](http://www.scb.se)

## ANNEX 1.8. CASE STUDY

### REMOTE DATA ACCESS FACILITIES - CANADA

#### 1. Broad description

Remote data access (RDA) is a mode of indirect access to confidential microdata through which researchers submit their own computer programs via the Internet to Statistics Canada, where they are run by Statistics Canada staff on the internal unscreened microdata. The results are then vetted for confidentiality and sent back to the researcher.

#### 2. Why is it good practice?

RDA fills a gap in the continuum of access to data. On the one hand, direct access to confidential microdata is restricted according to the provisions of the Statistics Act to employees of Statistics Canada and persons deemed to be employed under the Act (see Case Study – Research Data Centre Program – Canada). On the other hand, given limited resources, there can often only be a select number of products made available in an unrestricted manner from any given statistical activity. By having indirect access to the confidential microdata through the output of the computer programs that they submit, researchers from outside Statistics Canada can fulfil their own needs for tabulations or modelling while engaging relatively little of the agency's resources in the process. Agency staff vets the computer outputs before returning them to the researcher, thus ensuring data confidentiality.

#### 3. Target audience

RDA is available to all researchers who are not Statistics Canada employees and who have a demonstrated need to access microdata for statistical research. To prevent unnecessarily engaging the agency's resources, researchers must ensure that any products already in the public domain are insufficient to meet their needs.

#### 4. Detailed description

When using RDA, the researcher accesses the data through the output of a computer program that is executed by a Statistics Canada employee.

First, the researcher applies for RDA. At Statistics Canada, RDA is the responsibility of individual subject-matter divisions, and the service is managed at that level. Since no direct access to the microdata is involved on the part of the researcher, the approval process essentially ensures that the output will not be confidential and that information already in the public domain would not suffice to carry out the project.

Once a research project is approved, the subject-matter division provides the researcher the tools necessary to develop the programs before submission. The set of tools includes file names, record layouts and data dictionaries. In best-practice situations, a 'dummy' file is also made available by the subject-matter division. This is a data file with artificial data that mimics exactly the internal microdata, and which the researcher uses to develop and test the computer program prior to submitting it to Statistics Canada.

Once development and testing is complete, the researcher sends the program electronically, via e-mail, to the subject-matter division at Statistics Canada. The program is executed by survey staff using the internal microdata file. The program's output and log (containing diagnostics for the researcher to determine whether the program has executed properly) are vetted by agency staff to determine whether any confidential information is included. Any confidential information is deleted. If the amount of confidential information is large, the researcher may be asked to modify the program to reduce the output of confidential information, and to resubmit it. Then the output and log are sent electronically to the researcher.

Depending on the resources available to support RDA within the particular statistical activity, a small fee may be levied for use of the service. The fee, if any, is usually minimal compared to the costs that can be involved in requesting custom tabulations and/or analysis from the subject-matter area.

As a rule, the researcher is solely and fully responsible for the content and accuracy of the computer program. Arrangements can be made in certain cases, where agency staff will be called to participate in the development of the programs, and potentially in the analysis of the results. Such arrangements are negotiated in advance. Because they engage more Statistics Canada resources than basic RDA arrangements, extra fees are likely to be levied.

The vetting process can be time-consuming as it primarily involves manual work. To expedite this step, advance discussions with the researcher can indicate steps that can be taken with the program to reduce the time needed for vetting.

## **5. Supporting legislation**

Since researchers do not have direct access to confidential microdata, no specific legislative authority is invoked, apart from the Statistics Act which governs Statistics Canada in general and sets out the confidentiality requirements applied to all data prior to public release.

## **6. Strengths**

- Allows the use of unscreened microdata by researchers outside of Statistics Canada.
- Provides another mode of access to microdata, and thus is another means of expanding the outputs of the research community.
- Provides another opportunity for researchers to build on their capacity to work with microdata and enhance their analytical skills.
- Can be time-effective for smaller requests.
- Relatively inexpensive compared to other options for data access.

## **7. Weaknesses**

- Inconvenient to use in some ways, as the researcher does not see outputs prior to screening for confidentiality. This can make it more difficult to get a sense of small cell size and/or data accuracy.
- Not all software is supported. Researchers may have to learn new software or work with less familiar software.

- All output must be vetted for confidentiality prior to being returned to the researcher, engaging Statistics Canada resources.
- Requires that researchers learn and understand the content of the survey and microdata file, instead of relying on subject-matter staff as would be the case when requesting custom tabular and/or analytical output.

## 8. References

The various modes of access, including RDA, that are available for a number of surveys at Statistics Canada are well described in Tambay, J.-L., Goldmann, G., and White, P. (2001). *Providing Greater Access To Survey Data For Analysis At Statistics Canada*. Proceedings of the Annual Meeting of the American Statistical Association, August 5-9 2001.

More information can be obtained by contacting Statistics Canada staff or on the Statistics Canada web site ([www.statcan.ca](http://www.statcan.ca)). See [www.statcan.ca/cgi-bin/statcomment.pl](http://www.statcan.ca/cgi-bin/statcomment.pl)

## ANNEX 1.9. CASE STUDY

### REMOTE ACCESS FACILITY (FOR MICRODATA ACCESS) – AUSTRALIA

#### 1. Broad description

The Remote Access Data Laboratory (RADL) is a web-based tool that allows authorised users to access detailed microdata that is stored within the Australian Bureau of Statistics (ABS) secure environment. Built-in automatic checks prevent large-scale release of unit record information, thus maintaining confidentiality of data providers as outlined in Australian legislation.

#### 2. Why is it good practice?

The RADL provides access to more detailed and less confidentialised microdata than can be made available on CD-ROM. It provides greater flexibility in user analysis of microdata.

Access is limited to authorised users. All microdata remains within the ABS computing system. A balanced mix of automatic and manual processes prevent clients from obtaining outputs containing large amounts of unit record information. An audit trail is automatically maintained.

#### 3. Target audience

The RADL is primarily targeted at Australian government agencies involved in policy development and research areas within Australian universities. To a lesser extent, the RADL is also used for research purposes by the private sector and by non-profit institutions.

#### 4. Detailed description

Potential users of microdata are required to sign legal undertakings and read training material provided before RADL access will be granted. Authorised users are required to comply with published data-security guidelines and any further instructions of the ABS.

The RADL operates as a three-stage process. Clients submit batch-style queries via a secure section of the ABS website, which are firstly parsed for illegal commands. If the query is accepted, it is then executed in conjunction with ABS confidentialised microdata files. Finally, all produced output is automatically checked for confidentiality issues before being made available to clients on a secure web page.

A retrospective auditing process manually checks for inappropriate use of ABS microdata, and provides empirical evidence that automatic checks have been applied appropriately.

#### 5. Supporting legislation

The release of microdata by the ABS is governed by legislation, the Census and Statistics Act 1905. This legislation enables the Australian Statistician to release unit record data, provided this is done “in a manner that is not likely to enable the identification of a particular person or organization to which it relates”. Section 5 of Annex 1.1 provides more detail.

## 6. Strengths

- (i) Provides a secure online access point, from which users may access detailed ABS microdata from their own computing environments.
- (ii) Automatic protection of output at time of execution allows quick turnaround.
- (iii) Enables ABS to release more detailed microdata than that which can be released on CD-ROM.
- (iv) Flexibility of user analysis. Users are not restricted to a set of predefined tables.
- (v) Users are alleviated of CD-ROM security and data storage concerns.
- (vi) Statistical software is provided by ABS. Users do not need to supply their own licenses.

## 7. Weaknesses

- (i) Researchers still believe the conditions of release are too limiting and that the steps taken to make identification unlikely result in too little detail being released.
- (ii) Limited to batch-mode style of programming, lack of graphical-user interface functionality.
- (iii) Time taken to build automatic protections limits variety of statistical software packages made available.
- (iv) Heavy manual auditing load.

## 8. References

Australian Bureau of Statistics, (2005), *Responsible Access to ABS Confidentialised Unit Record Files (CURFs) Training manual*, Edition 2, Canberra, Australia, also available at [www.abs.gov.au->services we provide->curfs](http://www.abs.gov.au->services%20we%20provide->curfs).

Australian Bureau of Statistics, (2004), *The Remote Access Data Laboratory (RADL) User Guide*, Revised Version 2.0, Canberra, Australia, also available at [www.abs.gov.au->services we provide->curfs](http://www.abs.gov.au->services%20we%20provide->curfs).

Access to ABS CURFs web pages at [www.abs.gov.au->services we provide->curfs](http://www.abs.gov.au->services%20we%20provide->curfs).

## ANNEX 1.10. CASE STUDY REMOTE ACCESS TO MICRODATA FILES – DENMARK

### 1. Broad description

Statistics Denmark allows access to licensed (de-identified) microdata files for researchers and analysts. Access is only granted to employees (researchers and analysts) at institutions holding a special authorization issued by the General Director of Statistics Denmark. Special contracts are signed by the head of the institution and the researcher. Data are declared as confidential. Statistics Denmark has developed a remote access system allowing access to data from the researcher's own workplace.

Statistics Denmark does not give access to public-use microdata.

### 2. Why is it good practice?

The system allows access to very detailed de-identified microdata with a maximum of flexibility both to Statistics Denmark and the research environment. The system has replaced an on-site arrangement used for about 15 years. The researchers no longer have to work from premises in Statistics Denmark, which has allowed many more researchers to start research projects using microdata.

The technical system together with the authorization procedure and signed contracts is considered safe in relation to confidentiality. The basic microdata does not leave the premises of Statistics Denmark at any time, as only the statistical results are allowed to be transferred to the researcher.

The system is supported by the Danish Ministry of Research with a special grant. €800,000 per year is allocated to Statistics Denmark in order to reduce the costs of each project and with the vision that Danish researchers should develop to be among the best in the world to use register data.

### 3. Target audience

Authorizations can be granted to public research and analysts' environments (e.g. in universities, sector research institutes, ministries etc.) and to research organizations within a charitable organization.

Within the private sector, the following user groups can be granted authorisation if they have a stable research or analysts' environment (with a responsible manager and with a group of researchers/analysts):

- (i) Non-governmental organizations;
- (ii) Consultancy firms;
- (iii) Enterprises, although single enterprises cannot access microdata with enterprise data.

In order to grant an authorisation, Statistics Denmark will evaluate the proposed organization carefully, especially when it is an organization or firm within the private sector. Statistics Denmark will consider the credibility of the applicant in the light of ownership, educational standard among the staff and the research done for others.

Statistics Denmark will not grant authorization to single persons. Furthermore, media organizations are excluded from the scheme.

A ‘need to know’ principle is used as Statistics Denmark does not allow access to more data than needed according to the project description.

Researchers can have access to relevant business data after the “need to know” principle. Very few business data are excluded from remote access.

Only Danish research environments are granted authorisation as Statistics Denmark is not able effectively to enforce a contract abroad. Foreign researchers from well-established research centres can have access to Danish microdata from the on-site arrangement in Copenhagen or Århus. Visiting researchers can have remote access from a workplace in the Danish research institution during their stay in Denmark and under the Danish authorisation.

#### **4. Detailed description**

- (i) The scheme is administered centrally by the Division of Research Services. The staff of this unit also creates a substantial part of the inter-disciplinary data sets and have a general (authorized) access to all relevant Statistics Denmark data in order to reduce the administrative and bureaucratic workload. The scheme requires close cooperation between the Division of Research Services and the individual divisions. The advantage of such central organization is that the individual researcher is fully aware of who to negotiate with and who is responsible for the data set supplied.
- (ii) Statistics Denmark has not applied scrambling procedures or special grouping techniques to the data that are made available to the researchers. The data appear as in the basic register. It means that the linked data can be very detailed.
- (iii) The technical solution is web-based, as shown on the flow chart at the end of this case study.

The relevant microdata are produced by Statistics Denmark staff and the de-identified microdata are transferred to the disk storage connected to the special Unix servers. These Unix servers are only used by researchers and are separated from the production network.

Communication via the Internet is encrypted by means of a so-called RSA SecurID card, a component that secures Internet communications against unauthorised access. In practice the researcher rents a password key (a token) from Statistics Denmark. The token ensures that only the authorised person obtains access to the computer system.

A farm of Citrix Servers ensures that the researchers from their own workplace can ‘see’ the Unix environment in Statistics Denmark. All data processing is actually done at Statistics Denmark and data cannot be transferred from Statistics Denmark to the researcher’s computer. The researcher can work with the data quite freely and can make new data sets from the original data sets. The limit is of course the amount of disk space. Statistics Denmark has just increased the total amount of disk space considerably.

All results from the researchers' computer work can be stored in a special file and printouts are sent to the researchers by e-mail. This is a continuous process (every five minutes) and has proved to be quite effective. The advantage to Statistics Denmark is that all e-mails are logged at Statistics Denmark and checked by the Research Service Unit. If the unit finds printouts with too detailed data, the researcher is contacted to agree on details of the level of output. No severe violation of the rules established in the authorisation formula has so far taken place.

## **5. Supporting legislation**

Access to microdata is governed by the Danish Processing of Personal Data Act. The Act implements Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and the free movement of such data within the European Union. The previous act primarily governed registration and disclosure of data in registers, while the new Act applies to all forms of processing of personal data. The new term, "processing", covers all types of processing of personal data, including registration, storing, disclosure, merging, changes, deletion, and so on.

## **6. Strengths**

The remote access system together with the yearly grant from the Ministry of Research has increased the use of microdata for research significantly and has been evaluated as very satisfactory by the research community. From modest beginnings in 1986, the use of microdata has increased markedly for researchers at Statistics Denmark. In 1997, 71 researchers used the on-site arrangement, while in 2005 under the scheme for remote access through the Internet the figure rose to more than 300. 132 environments had been granted authorization by August 2005.

## **7. Weaknesses**

The remote access system is from time to time under heavy pressure from an increasing number of users. The need for continuous upgrading of the computers and disk space is sometimes difficult to finance.

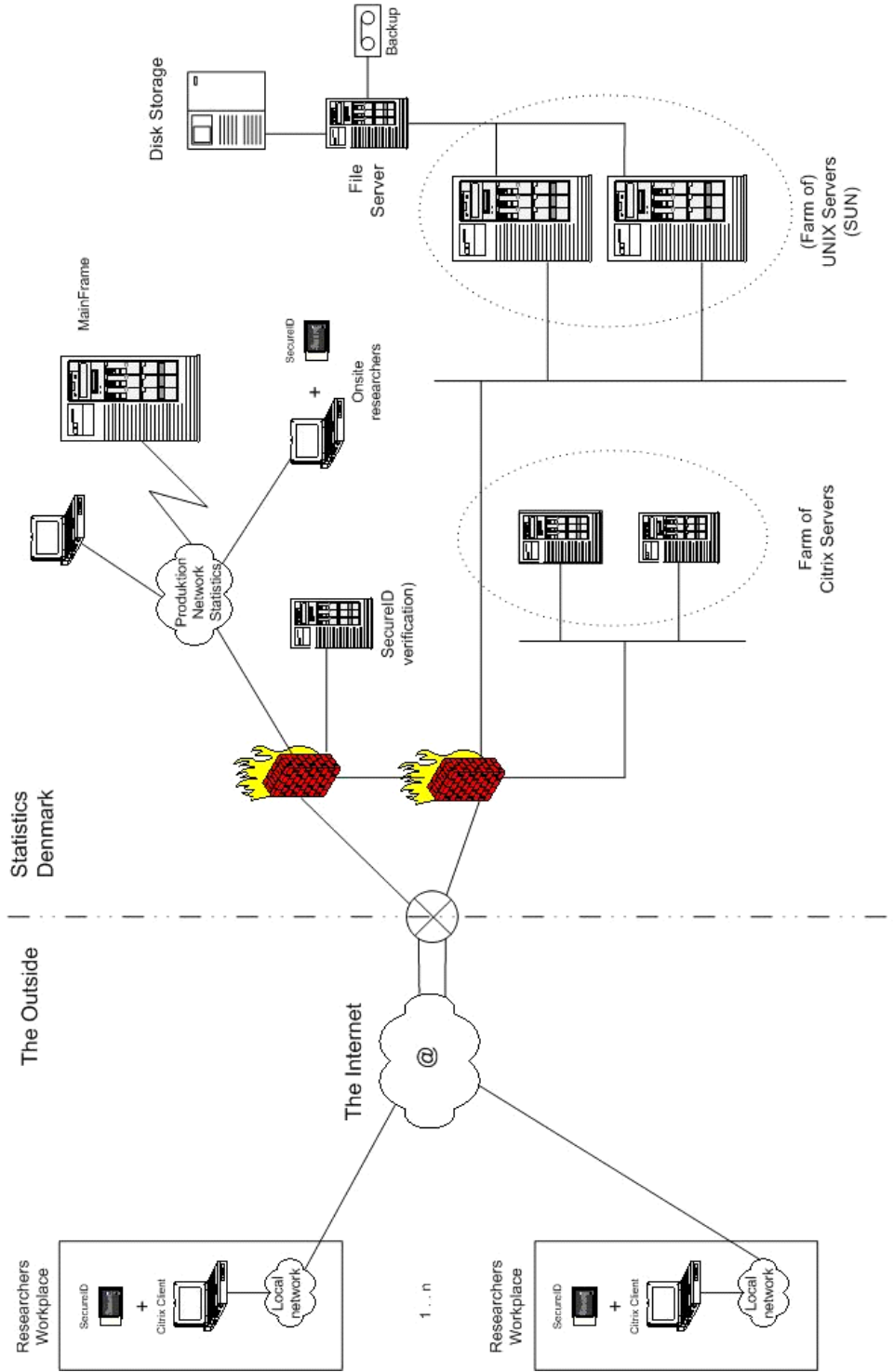
Researchers are still unsatisfied with the costs. Although access to the remote access system is generally free of charge, the researchers have to pay (by the hour) for the creation of the data sets.

## **8. References**

Otto Andersen: *From on-site to remote data access*, contributed paper to the Joint ECE/Eurostat work session on statistical data confidentiality (Luxembourg, 7-9 April 2003).

Overleaf is a scheme showing how remote access to Statistics Denmark microdates operates.

Remote Access to Statistics Denmark. January 2003. Principles of Operation



Rev 1 March 13th 2002  
ABJ/

## ANNEX 1.11. CASE STUDY RESEARCH DATA CENTRE PROGRAM – CANADA

### 1. Broad description

Starting in 2000, Statistics Canada, in partnership with participating Canadian universities, the Social Sciences and Humanities Research Council and the Canadian Foundation for Innovation, established a network of Research Data Centres (RDC) in Canadian universities. These centres are enclaves of Statistics Canada, within which researchers have access to household survey data in an environment that respects Statistics Canada's requirements for security and confidentiality. There are currently 15 RDC locations across the country, plus a federal RDC in Ottawa used by statistical researchers in federal government departments.

### 2. Why is it good practice?

The Chief Statistician and the President of the Social Sciences and Humanities Research Council (a granting council for the human sciences) established a panel of highly qualified individuals to assess how social sciences could become more relevant and more quantitatively oriented. The panel observed that Canada's capacity in the quantitative social sciences was stagnating, due in part to difficulties in accessing the data needed to conduct analyses on some of the important socio-economic and demographic issues facing Canadian society. It was also observed that the advent of complex longitudinal survey data made it very difficult (if not impossible) to create useful public-use microdata files. The Research Data Centre Program successfully addresses the issues of access while respecting the provisions of the Statistics Act for security and confidentiality of the data. The RDC Program makes it possible for researchers outside Statistics Canada to directly access microdata (after satisfying several conditions) that would otherwise not be available to them. Society benefits because insights are gained that would otherwise not be possible; and the statistical system benefits because the visible relevance of available statistical information increases. The following are the basic features:

- Academics wishing to access confidential microdata submit a research proposal which is peer reviewed by the Social Sciences and Humanities Research Council;
- Authors of accepted proposals are sworn in under the Statistics Act as employees of Statistics Canada, subject to all the conditions and penalties of the Statistics Act;
- All work is carried out in an RDC, which in turn is supervised by a regular Statistics Canada employee;
- Academics must submit a short article to SC for potential publication.

The proposal is based on two features of the Statistics Act: first, that it explicitly mandates Statistics Canada to analyse data; and second, it allows the agency to swear in under the Act personnel needed to carry out its mandate.

### 3. Target audience

The RDC Programme is open to all researchers who are not employees of Statistics Canada and who require microdata for statistical research. This includes established academics, new researchers, graduate students and researchers from federal departments and provincial governments.

### 4. Detailed description

A researcher (or research team) seeking access to the detailed microdata must submit a proposal that outlines the analyses to be conducted. The proposal must be a maximum of five pages excluding the CV of the researcher(s) and it must contain the following information:

- Project title;
- Rationale and objectives of the study, including: specific questions or objectives of the project; and how the research will contribute to the knowledge in the field of study;
- Proposed data analysis and software requirements including: the proposed statistical methodology; its suitability for this project; and the software needed;
- Data requirements including: why access to the confidential data (as opposed to public use microdata files) is necessary; the survey file/files or cycles to be used; a description of the specific population of interest; and a list of the variables to be used;
- Expected project start and end dates; and
- References – sources of quotes used in the proposal or for specific analytical methods employed.

All proposed research projects are reviewed by a peer-group committee who determine the academic merit of the work and the suitability of the methods and the data. They also verify that the work can only be undertaken with access to the confidential data files. In cases where the Public Use Files would be suitable or the work lacks rigour or focus, the application will be denied. Approved researchers have access to the required files within a RDC, but only results that have been screened for disclosure protection can be removed from the RDC. Researchers are required to produce a report for Statistics Canada as part of their commitment under the Statistics Act (the so-called “deemed employee” provisions of the Act). Once that obligation is fulfilled, researchers are free to publish other articles that may be based on the research project.

Before being granted access to the data, researchers must undergo a security check; sign the oath/affirmation of secrecy required by the Statistics Act; acknowledge in writing that they have read and understood the relevant sections in the Statistics Act and specifically the policies related to data confidentiality and security; acknowledge in writing that they have read the documentation on conflict of interest and declaring that they will comply with the requirements.

The RDC network has substantially increased the access by researchers to the complex detailed microdata survey files. As of June 2005 there were over 500 active projects and over 1,300 researchers in the centres. Approximately one third of the researchers are students. There are also over 280 articles, book chapters, working papers and theses that have been published from the research conducted in the centres.

## 5. Supporting legislation

Section 5 of the *Statistics Act* permits persons carrying out any function or performing work for Statistics Canada to become “deemed employees”, thereby allowing access to the confidential data files.

## 6. Strengths

- Allows access to data outside the Statistics Canada offices while continuing to respect the requirements of the Statistics Act.
- Increases the opportunity to conduct research on key socio-economic and demographic issues and expands the quantity and range of research outputs using statistical data.
- Effective in developing the next generation of quantitative social scientists in Canada.
- Provides greater research opportunities for highly qualified analysts who do not reside in cities in which Statistics Canada has offices.
- As the use of the data increases, greater feedback is obtained on the surveys and the data sets that they generate. This results in quality improvements in the data and it opens new possibilities for the use of the data.
- Accelerates the development of advanced statistical methods required to analyse complex survey data.
- Provides Statistics Canada with much more detailed and timely information on the use of its data.
- Locating RDCs in universities reduces the cost of conducting research since it eliminates the need for travel for many researchers.
- The research conducted in the RDCs adds substantially to the body of literature on major social, economic and demographic questions affecting Canadian society. It also serves to inform public policy and debate.

## 7. Weaknesses

- RDCs are costly to build, manage and operate. This places them out of the reach of some of the smaller universities in Canada.
- All output must be vetted for confidentiality prior to leaving the RDC. This is a manual effort and, even when prompt attention is given to the vetting, results in some delays to the researcher.
- To date, data sets in the RDCs have been from household surveys. Although the demand exists and there are no technological barriers, access to data from the census of population and from business surveys are not placed in the RDCs.

## 8. References

Statistics Canada’s Policy on the Use of Deemed Employees is available on request. More information on the Research Data Centre Program can be viewed at the Statistics Canada web site at: [www.statcan.ca/english/rdc](http://www.statcan.ca/english/rdc).

## ANNEX 1.12. CASE STUDY RESEARCH DATA CENTRES – UNITED STATES

### 1. Broad description

Research Data Centres (RDCs) offer qualified researchers restricted access to confidential economic and demographic data collected by the U.S. Census Bureau in its surveys and censuses. All projects must offer benefits to U.S. Census Bureau programmes. These projects are carried out at U.S. Census Bureau headquarters, or at one of eight other secure locations around the U.S.

### 2. Why is it good practice?

The statutory provisions under which the U.S. Census Bureau collects data prevent the release of the full details of survey data (e.g. names, addresses) in order to protect the confidentiality of respondents. The microdata provided by businesses are never released to the public; public use microdata samples of household surveys include limitations on geography, topcodes on income, collapsing of occupational categories, and so forth. Nevertheless, some research would benefit from access to this additional information. A ‘research enclave’ where data dissemination is tightly controlled allows the estimation of statistical models based on the full data set.

### 3. Target audience

RDCs are aimed at researchers in academia; at independent research organizations such as the National Bureau of Economic Research; and in federal, state, and local government agencies. Tabulations of confidential data are generally not allowed to be removed from the RDCs, and therefore estimation of statistical models is the focus of their activities. All researchers are required to become Special Sworn Status employees of the U.S. Census Bureau, and as such are subject to the penalty provisions of its authorizing legislation (e.g. a fine of US\$250,000), should there be a confidentiality violation.

### 4. Detailed description

The objective of the U.S. Census Bureau and the RDCs is to increase the utility and quality of U.S. Census Bureau data products. Access to microdata encourages knowledgeable researchers to become familiar with U.S. Census Bureau data products and data collection methods. More importantly, providing qualified researchers access to confidential microdata enables research projects that would not be possible without access to respondent-level information. This increases the value of data that has already been collected. Access to the microdata also allows for data linking that is not possible with aggregates – both cross-survey linkages and longitudinal linkages. These linkages leverage the value of existing data. Creative use of microdata can address important policy questions without the need for additional data collection.

In addition, the best means by which the U.S. Census Bureau can check the quality of the data it collects, edits, and tabulates is to make its microdata records available in a controlled, secure environment to sophisticated users who, by employing the microdata records in the course of rigorous analysis, will uncover the strengths and weaknesses of those records. Each set of observations is the end result of many decision rules covering definitions, classifications, coding procedures, processing rules, editing rules, disclosure rules, and so forth. The validity and consequences of all these decision rules only become evident when the U.S. Census Bureau's micro databases are tested in the course of analysis. Exposing the conceptual and processing assumptions that are embedded in the U.S. Census Bureau's micro databases to the light of research constitutes a core element in the U.S. Census Bureau's commitment to quality.

The opportunities for researchers to carry out unique research come at a price. Research conducted at RDCs takes place under a set of rules and limitations that are considerably more constraining than those prevailing in typical research environments. The process is described below.

Working closely with an RDC administrator, researchers develop a preliminary research proposal that includes information about the researcher(s), site where the research will be carried out, its purpose, funding source, requested data sets, desired software, a brief narrative description of the research project and proposed benefits to the U.S. Census Bureau. The researcher enters this information via the online proposal management system. Once a preliminary proposal has been submitted, the RDC administrator reviews it and advises the researcher of any suggestions for improvement or refinement. The administrator must approve the preliminary proposal before the researcher can submit a final proposal to the U.S. Census Bureau's Center for Economic Studies (CES) for final review.

Research proposals submitted to CES are reviewed on the basis of five major criteria:

- **Benefit to U.S. Census Bureau programmes.** Proposals must demonstrate that the research is likely to provide one or more benefits to the U.S. Census Bureau. These benefits can include:
  - Understanding and/or improving the quality of data produced through a Title 13, Chapter 5 survey, census, or estimate [Title 13 is the U.S. Census Bureau's authorizing legislation];
  - Leading to new or improved methodology to collect, measure, or tabulate a Title 13, Chapter 5 survey, census, or estimate;
  - Enhancing the data collected in a Title 13, Chapter 5 survey or census, for example
  - Improving imputations for non-response, or developing links across time or entities for data gathered in censuses and surveys authorized by Title 13, Chapter 5;
  - Identifying the limitations of, or improving, the underlying Business Register, Household Master Address File, and industrial and geographical classification schemes used to collect the data;
  - Identifying shortcomings of current data, collection programmes and/or documenting new data collection needs;
  - Constructing, verifying, or improving the sampling frame for a census or survey authorized under Title 13, Chapter 5;

- Preparing estimates of population and characteristics of population as authorized under Title 13, Chapter 5;
  - Developing a methodology for estimating non-response to a census or survey authorized under Title 13, Chapter 5; and
  - Developing statistical weights for a survey authorized under Title 13, Chapter 5.
- **Scientific merit.** This criterion relates to the project’s likelihood of contributing to existing knowledge. Evidence that a federal agency such as the National Science Foundation or the National Institutes of Health has approved the proposal for support constitutes one indication of scientific merit.
  - **Clear need for non-public data.** The proposal should demonstrate the need for and importance of non-public data. The proposal should explain why publicly available data sources are not sufficient to meet the proposal’s objectives.
  - **Feasibility.** The proposal must show that the research can be conducted successfully with the methodology and requested data.
  - **Risk of disclosure.** Output from all research projects must undergo and pass disclosure review.
    - Tabular and graphical output presents a higher risk to disclosure of confidential information than do coefficients from statistical models.
    - The U.S. Census Bureau is required by law to protect the confidentiality of data collected under its authorizing legislation.
    - Some data files are collected under the sponsorship of other agencies. In providing restricted access to these data, the U.S. Census Bureau Center for Economic Studies (CES) must adhere to all applicable laws and regulations.
    - Researchers may be required to sign non-disclosure documents of survey sponsors or other agencies that provide data for their research projects.

Both U.S. Census Bureau and external experts on subject matter, data sets and disclosure risk review all proposals. Relevant data sponsors and data custodians also review proposals that request certain data sets. Any proposals seeking to use data sets that contain Federal Tax Information must also be reviewed for approval by the Internal Revenue Service.

All of the actual processing of data for approved proposals is conducted on servers located in the U.S. Census Bureau’s secure central computer facility. Researchers located in the RDCs use ‘thin client’ terminals to access these servers via encrypted communication lines.

## 5. Supporting legislation

Title 13, United States Code, permits the U.S. Census Bureau to employ Special Sworn Status employees for the purpose of carrying out its mission. Specifically, Section 23(c) states:

“The Secretary [of Commerce] may utilize temporary staff, including employees of Federal, State, or local agencies or instrumentalities, and employees of private organizations to assist the Bureau in performing the work authorized by this title, but only if such temporary staff is sworn to observe the limitations imposed by section 9 of this title.”

## 6. Strengths

- (i) As administrative data about individuals becomes more and more available through the Internet, statistical agencies must reduce the detail about individuals available through public use microdata. The availability of such data through the RDCs as research enclaves can help ensure that valuable research can continue.
- (ii) Since business microdata has never been in the public domain, the RDCs allow microeconomic research on businesses that could not otherwise take place.
- (iii) There is potential for expansion to allow the confidential data of other federal agencies to be available through the RDCs.

## 7. Weaknesses

- (i) Operating the RDCs has costs, some of which must be absorbed by the U.S. Census Bureau.
- (ii) The proposal review process is cumbersome and time consuming, and the consequent delays in getting access to the data at the RDCs are frustrating to researchers.
- (iii) All projects must, by law, have a benefit to the U.S. Census Bureau. Therefore, some worthy research projects with questionable benefits must be rejected.

## 8. References

The CES web site contains additional information about the RDC programme: <http://www.ces.census.gov/ces.php/rdc#objectives>.

Prepared by Dr. Daniel Weinberg, Chief, Center for Economic Studies, and Chief Economist, U.S. Census Bureau, October 13, 2005.

## ANNEX 1.13. CASE STUDY DATA LABORATORY ARRANGEMENTS – NETHERLANDS

### 1. Broad description

After the initial success of releasing licensed microdata files (anonymised and protected against disclosure) from its social sample surveys from 1994 onwards, researchers also developed a demand for accessing other microdata files: business survey microdata, fiscal income data, cause-of-death data, and the like. Such data are much less easy to protect against disclosure because important variables are very skewed in their distributions, samples are much more stratified and even integral in some (or all) strata, and so on. The solution to serve the needs of these researchers was to create a separate research facility, a data laboratory, within the safe walls of Statistics Netherlands' (Central Bureau voor de Statistiek, or CBS) two establishments, with most universities within an hour's reach. So with this solution the microdata does not go to the researcher but the researcher comes to the microdata. In 1998 the Centre for Research on Economic Microdata (Cerem) was established after fairly long consultations with business representatives who needed to be convinced of the utility and the safety of academic on-site analysis of business microdata. Subsequently on-site facilities were also used by researchers on social microdata with more detail than can be found in the licensed microdata files, and by researchers on social microdata from matched administrative data.

More recently the Centre for Policy related Statistics (CPS) was set up as a department of Statistics Netherlands in 2002 in response to an increasing demand for statistical information by government policymakers and national planning agencies. One important factor in this respect was the demand for statistics to measure the effects of policy measures, and to gain insight into possible effects of a change in policy measures.

These demands are mostly of a short-term character. Almost inevitably the statistical programme of a national statistical agency like Statistics Netherlands (SN) is not suited for short-term changes. Instead the programme is based on producing statistics that can be compared over time and therefore have a slow rate of change. This makes it difficult to fulfil the needs of government departments and planning offices in this respect.

The CPS was set up to improve the fulfilment of these demands. Flexibility was achieved by making the working programme of the Centre dependent on the demands of the departments and planning offices. Only a few of the staff are paid directly from the budget of SN. Other demands for statistics are paid directly by the departments. A pilot project was set up with the Ministry of Social Affairs and Employment in 2002. This pilot was successful and led to a further increase in services.

## 2. Why is it good practice?

The data laboratory arrangement makes it possible to have microdata analysed in a safe setting. The microdata themselves cannot always be protected; for example, the producer of light bulbs from Eindhoven will always be recognisable and indispensable at the same time. But the settings in which these microdata are analysed are fully controlled. The number of days spent in the data laboratories is increasing each year by more than 20% and has now surpassed 1500.

## 3. Target audience

Microdata are made accessible under a contract or license to legitimate researchers only. Section 41 of the law cites the researchers that are qualified. These include the universities and other research institutes with a legal foundation, but also Eurostat and the EU NSOs. A residual category of applicants must be formally admitted by the Central Commission for Statistics (CCS), the supervisory body for CBS. The CCS has set its own criteria and procedures for deciding, in which a focus on statistical (aggregate) research, independence from administrative authorities, and the intention to share results in the public domain are predominant. The CCS has no principal objection against admitting non-EU universities, for example, but a commercial bank or a journalist would not be eligible. In practice, researchers seem to have better methodological qualifications when working on site, if only because their microdata and statistical models and software are complex in comparison with the social analysis of official social sample survey microdata at the universities. At present only researchers affiliated with Dutch research institutes are allowed.

## 4. Detailed description

The services of CPS can be divided into three groups.

- First there is the advisory function. The Centre can give advice about the possibilities of doing research on a specific question. Because the CPS staff have a broad knowledge of the available data sources, this can help to reduce data collection costs.
- Secondly, as mentioned above, CPS can itself conduct research on request. This research is done solely on existing data material; no additional survey activity is undertaken in this respect.
- A third activity is the possibility to provide access to microdata for researchers from outside SN. The microdata are made available at the level of individual records where, of course, direct identifiers are removed.

The datasets are all well documented, making it possible for researchers to gain access to the material efficiently and to evaluate the relevance of the dataset for their research.

Furthermore, precautions are taken to ensure that the security of the process is maximised.

Most microdata are made available for analysis on-site: i.e. researchers work at one of the offices of Statistics Netherlands on a dedicated infrastructure. For reasons of security, this infrastructure is physically disconnected from Statistics Netherlands' production environment and visitors only have access to the micro datasets they need for their specific research.

A second type of microdata service is *remote execution*. Using this service, researchers may send in scripts to be executed on well-defined sets of micro data.

For all types of microdata services, checks on the possibility of statistical disclosure are performed before results are made available to researchers for use outside the secure environment.

From mid-2005 a *remote access* facility has been developed, making it possible for researchers to analyse microdata present at SN through a secure connection from workstations in their own institute. This facility is much like the facility in use in Denmark.

Part of the security regime here is the use of a secure Internet connexion and the application of biometric identification (fingerprints) and PKI certificates.

The on-site facilities for microdata analysis within SN developed in recent years originate from different departments, and were designed slightly differently. Apart from the on-site facility within the CPS, there were separate facilities within the department for social statistics and a facility for economic statistics, known as CEREM (Centre for Research on Economic Microdata). At the end of 2005 all the activities for microdata services were pooled within the CPS, enabling a more transparent and more efficient microdata access. This means that, although CEREM as such is no longer in existence, CPS offers the same facilities and availability for microdata analyses.

The use of microdata is expected to grow rapidly the coming years. In this respect, the remote access facilities in particular are promising.

## **5. Supporting legislation**

In 2003 the statistical legislation allowing the release of microdata was rephrased to make formally possible the analysis of microdata on site in a data laboratory. Section 41 now makes it possible "to provide or grant access to a set of data".

## **6. Strengths**

The main strength from the perspective of the researcher is that there is hardly a limit to the amount and nature of microdata that can be analysed.

## **7. Weaknesses**

The main remaining weakness, at least from an international perspective, is the language used for documentation. Plans for making the documentation in English are foreseen for the coming years.

## 8. References

The Centre for Policy related Statistics can be reached by e-mail at [CvB@CBS.nl](mailto:CvB@CBS.nl). Within the CPS, contact persons are Frans Hoeve (+31 70 337 5609 or [FHVE@CBS.nl](mailto:FHVE@CBS.nl)) and Gerhard Meinen (+31 70 337 4228 or [GMNE@CBS.nl](mailto:GMNE@CBS.nl)).

## ANNEX 1.14. CASE STUDY DATA LABORATORY MICRODATA ACCESS - NEW ZEALAND

### 1. Broad description

Access to anonymised unit record data is provided to researchers in a secure environment within the regional offices of Statistics New Zealand (SNZ).

Access to microdata is governed by the provisions of the New Zealand Statistics Act 1975, and is implemented in accordance with SNZ's microdata access protocols (see [www.stats.govt.nz/about-us/policies-and-guidelines/general/microdata-access-protocols.htm](http://www.stats.govt.nz/about-us/policies-and-guidelines/general/microdata-access-protocols.htm)). All microdata access requests are subject to the approval of the Government Statistician.

Microdata access can be supplied to New Zealand government departments for bona fide research or statistical purposes, or to researchers contracted to SNZ (who must provide some output from their work that is of direct value to the Official Statistics System), and to other government agencies and bodies in New Zealand when data has been collected jointly. The Government Statistician can also permit access to microdata when written consent has been obtained from all the people who supplied the data.

To request access to unit record data through the data laboratory, a researcher must complete an application form which includes adequate detail on the nature of the intended research, what variables are required, and the proposed outcomes of the research, such as publications, presentations, or a contribution to ongoing research. An initial application form is completed by the applicant. This is assessed by relevant staff of SNZ, who will work with the applicant to produce a final application that is feasible and complies with the department's microdata access criteria. Once an application is in a form which meets the department's criteria, it is submitted to the Government Statistician, who makes a decision to approve or refuse the request. If an application is refused on some grounds, the applicant can address that issue and resubmit an altered application.

### 2. Why it is good practice?

The data laboratory balances the need to ensure that New Zealanders have confidence in SNZ's ability to protect their identities and personal information with the value this information has for conducting research and developing Government policy.

The data laboratory gives sophisticated researchers access to unit record data that is not otherwise available. The rigorous process for approving applications ensures that the provisions of the Statistics Act 1975 are always taken into account when access to unit record data is granted, and that access to the unit record data is necessary for the proposed work.

A number of techniques are used to limit the disclosure risk. Unit record data are anonymised and modified to protect respondent identities. Data sets are made available in a secure physical and computing environment to prevent unauthorised access to the data. Statistical outputs generated in the data laboratory are checked to guard against disclosure risks. Finally, all papers and reports produced based on data laboratory research are checked prior to publication.

### 3. Target audience

The data laboratory is aimed at researchers and analysts working in New Zealand. In some instances a visiting foreign academic, who can point to benefits to New Zealand's Official Statistical System from their study, may be given consideration.

### 4. Detailed description

The following steps outline the assessment and approval process that follows the receipt of an initial data laboratory application form. During this process, staff at SNZ and the researchers seeking access to microdata may be involved in frequent communication. The data laboratory administrator coordinates communication and ensures that agreements and decisions are recorded.

1. The subject-matter area (SMA) unit responsible for the data to which access is requested examines the proposal in terms of the fitness for purpose of the proposed data set and variables requested for the study. Depending on the basis for permitting access to the microdata relevant issues may include sample sizes, data quality, confidentiality, and whether the research will benefit the official statistical system. SMA staff provide an assessment, and the manager of the SMA unit then makes a recommendation on whether the requested access to microdata should be approved or not.
2. The Statistical Methods unit examines the proposal for potential breaches of confidentiality, and specifies modifications to the data (typically removal or aggregation of geographic variables and randomisation of IDs) to minimise these risks. The manager of the Statistical Methods unit then makes a recommendation on whether the requested access to microdata should be approved or not.
3. A staff member from Strategic and Financial Services determines if the application is consistent with the requirements of the Statistics Act 1975 and SNZ's microdata access protocols.
4. The Microdata Access Manager prepares a summary of the issues raised for consideration by senior management.
5. The Group Manager responsible for the SMA provides their comments, and a recommendation to the Government Statistician to approve or reject the requested access to microdata.
6. The Government Statistician approves or rejects the application for microdata access through the data laboratory. The researcher is then advised of the decision. If the application is refused, the researcher will be notified of the reasons for the refusal. The researcher can subsequently resubmit an application. This is usually done after modifying the proposal to address the particular grounds for refusal.

If the proposal is accepted, the SMA unit creates a customised data set designed to provide only the information needed for the research. Statistical Methods staff check this data set before it is copied to the data laboratory. A contract is drafted and signed prior to researchers beginning their work. The restrictions on how data may be used are set out in the contract, and these obligations are made clear to each user who is given access to data. All researchers must also sign the statutory declaration of secrecy, required under the Statistics Act 1975, before beginning to work with the data. Once an agreement has been signed, changes such as the addition of new researchers to the agreement, are managed by way of letter of variation. Signatory rights are limited to Deputy Government Statistician level.

7. While research is underway, Statistical Methods staff check all outputs to ensure that there are no confidentiality breaches. All draft publications are also submitted to SNZ for checking.

## 5. Supporting legislation

The release of microdata into the data laboratory environment is governed by the Statistics Act 1975.

## 6. Strengths

- (i) The data laboratory allows for access to the most detailed data available for users working within the secure environment.
- (ii) Contracts provide a legal framework to ensure confidentiality protection for these data sets.
- (iii) Sanctions can be applied to users and organizations that breach the agreements, and this helps to ensure use of data sets is appropriate.

## 7. Weaknesses

- (i) Researchers complain that the timeframe for approval is too lengthy. SMAs have found that it can be difficult to balance their regular work programme with the uneven demands placed on them by data laboratory projects.
- (ii) The recovery of costs for access, support and initial data set development is problematic for some academic researchers.
- (iii) The care and attention given to the approval of proposals and the checking of outputs can be expensive and time-consuming.

## 8. References

Datalab: [www.stats.govt.nz/products-and-services/datalab.htm](http://www.stats.govt.nz/products-and-services/datalab.htm)

Statistics Act 1975: [www.legislation.govt.nz/](http://www.legislation.govt.nz/)

Microdata access protocols:

[www.stats.govt.nz/about-us/policies-and-guidelines/general/microdata-access-protocols.htm](http://www.stats.govt.nz/about-us/policies-and-guidelines/general/microdata-access-protocols.htm)

## ANNEX 1.15. CASE STUDY DATA LABORATORY MICRODATA ACCESS - BRAZIL

### 1. Broad description

Statistical dissemination in the Brazilian Institute of Geography and Statistics (IBGE) was traditionally carried out in two ways: for the general public, by means of media communication, assisted by media releases or press conferences; and for the general users, through printed publications and electronic publications. For more specialized users and government agencies, the requirements are met through customized tables and public use microdata files.

A policy of free dissemination of all products through Internet has been adopted in IBGE, since 2001. There has been outstanding growth in this communication channel.

As well as the electronic publications, the IBGE web page contains two important databases: Aggregated Statistical Tables (SIDRA) – a database with information grouped at territorial level that allows the users to construct tables according to selected information; and Multidimensional Statistical Database (BME) – a database with microdata information that allows users to construct tables according to selected information and confidentiality constraints. This database requires Internet subscription.

IBGE has been releasing public use microdata files for households' statistics since the early 1990s. Measures taken to protect the confidentiality of these microdata include suppression of geographical detail. However, no public use microdata files are released for businesses data, or for the 1996 Agricultural Census and the short form 2000 Population Census.

The pressure of increasing demand, the advance of technology and the increase of sensitivity to privacy issues have encouraged the development of arrangements to provide restricted access for researchers to data files that the statistical agency does not release to the general public. These arrangements permit a more in-depth analysis than was possible when using tabular aggregated data. This is done in IBGE via on-site access at the headquarters of the agency.

This Case Study provides short summaries of the procedures that have been implemented and are currently in use by IBGE since 2003, in order to permit external researchers, analysts within government, academia and other organizations to access restricted data.

### 2. Why is it good practice?

Confidentiality is a key element of respondents' trust, thus maintaining their cooperation in the provision of accurate data. As a result, the policy for the release of data is to prevent disclosure of information about individual persons or businesses, consistent with IBGE's legislation supporting confidentiality.

But it is also essential to try to reach the needs of the research community while maintaining confidentiality and security.

To provide restricted data access for analysis requires collaboration between all involved parties and preparation to deal with a variety of situations and questions. Technical developments may allow for new ways of achieving the needs of the research community whilst maintaining confidentiality and security.

### 3. Target audience

The target audience is researchers requiring special data access to information not available through the web site or public use data files.

### 4. Detailed description

The following describes the administrative and technical measures to regulate the access of restricted microdata and to ensure that the output is released with an adequate level of protection so that individual data cannot be disclosed. The procedures cover the following steps:

(1) - application

The researcher submits the research project to be evaluated if it is for public or academic interest, for statistical purposes and also whether it is feasible.

(2) - evaluation of the project

A Committee of Assessment of Restricted Data Access evaluates the project, based on submissions of the thematic area responsible for the survey microdata. The Committee authorizes (or not) the access to internal data files under the appropriate conditions.

The Committee is chaired by the Deputy Director for Surveys and composed of senior staff members dealing with business, methodology and dissemination coordination.

(3) - formal agreements to access

Once a project has been authorized, formal agreements between the researcher and the agency are established. These agreements involve a written contract (contractual arrangement), and an agreement form outlining the conditions of access and setting out fees for the proposed work.

(4) - on-site access

The databases are installed in the room with special computers for the researchers. The security features of the computers include a blockade to external networks to prevent transfer of data. Furthermore, the external disk drives and serial parallel ports are disabled. The identification of the enterprises is recoded in the databases from businesses surveys of IBGE or from external sources.

The researchers do the work and save the output in the hard disk of the special computer and then prepare a report document. A CD-Rom with this information is prepared by IBGE staff, to be analysed by the thematic survey area.

## (5) - evaluation of output

The statistical output must be analysed before its release to the researcher to ensure the technical assessment of disclosure risks and confidentiality requirements. The analysis is undertaken by the thematic area responsible for the survey microdata, the same that gave submissions for the committee decision.

## (6) - releasing the output

Once the output of the project has been approved, i.e. the thematic area judges that there is no risk of disclosure, another formal agreement is established. This new agreement outlines the conditions of use of the data generated by the special access, i.e. the user has to recognize that data are the property of IBGE and has to provide advice of this special access when releasing the results and analysis involving these data.

Table 1 shows the number of projects analysed by the Committee from September 2003 to February 2006. In 37 projects analysed, 3 projects involved data from the long form of the 2000 Population Census. In this case, the researcher needed different geographical areas from the weighting areas used in the sampling weighting process. One project involved data from an annual trade survey; one from an annual services survey; 30 projects from manufacturing surveys; and 2 projects involved data from manufacturing, trade or services surveys simultaneously.

**Table 1** – Number of Projects Analysed by the Committee  
(September 2003 – February 2006)

Thematic area	Number of projects
<b>Total</b>	<b>37</b>
Population Census	3
Trade Survey	1
Services Survey	1
At least 2 businesses surveys	2
Manufacturing surveys	30

## 5. Supporting legislation

The regulations for the provision of restricted data access were established by IBGE using the following expedients:

- Resolution of the Board of Directors, n. 7, of May 29, 2003 – that created the Committee of Assessment of Restricted Data Access.
- Regulation of the Chief Statistician, n. 485, of July 8, 2003 – that appointed the members of the Committee.

- Regulation of the General Coordinator of the Centre for Documentation and Dissemination of Information, n.1, of September 10, 2003 – that established the objectives of the rooms for use in the on-site restricted access.

## 6. Strengths

Provides a secure way of providing researcher access to IBGE data for projects that are of clear statistical or academic benefit.

## 7. Weaknesses

Although about 40 projects have been working on this on-site system at IBGE since 2003, we have had a lot of difficult tasks to face. It has been:

- time-consuming to analyse projects because, in many cases, there is a need to contact the proponents to redesign the project or to provide detailed explanations of why the project is not feasible;
- time-consuming to prepare user-friendly documentation;
- time-consuming to analyse the outputs due to faults in the documentation.

In general, the expected work time is underestimated.

Another issue involves managing the tension between the agency and the researchers in regard to the acceptability of the current practice. The culture and value system of the research community is very different from that of a National Statistical Office.

Researchers still think of microdata access arrangements as unnecessary bureaucracy, too limiting and inconvenient. This lack of convenience for the researcher includes the requirement to work at the agency. That can be an expensive option, especially for researchers living in other cities or countries. Another point is that sometimes the researcher is forced to use unfamiliar data analysis software.

There is an internal debate about the acceptability of this practice. Even under measures to regulate the access of restricted microdata, there is a worry that it could still alarm public opinion with suspicion of disclosure. The reaction of respondents would have some impact on response rates.

Increasingly, researchers are looking to link data sets with the data sets of the agency. Although matching of databases brings benefits, the identification risks increase.

There are some issues concerning transparency. The IBGE web site was an effective way to provide information on how to make access available for researchers. However, information about the procedures is only provided through Intranet and the users learn about the procedures only when asking for special data.

Therefore, it is a challenge for us to be transparent about the arrangements of providing access to data for researchers under controlled conditions for specific purpose. But the visibility of such arrangements is necessary to increase public confidence that microdata will be used properly. We would want to be completely transparent about the specific uses of microdata to avoid suspicion of misuse and ensure that researchers are aware of the consequences for them and their institution if there are breaches of confidentiality. On the other hand, there is a fear of excessively increasing the demand.

There is a demand to install rooms for on-site access outside the headquarters of the agency, especially in the big cities like São Paulo and Brasília. But to meet this demand requires investment in resources to train staff and prepare the infrastructure.

## 8. References

IBGE (2003), Resolução do Conselho Diretor nº 7, de 29.05.2003. (Resolution of the Board of Directors of IBGE, n. 7, of May 29, 2003 – that created the Committee of Assessment of Restricted Data Access).

IBGE (2003), Portaria do Presidente nº 485, de 08.07.2003. (Regulation of the IBGE's Chief Statistician, n. 485, of July 8, 2003 – that appointed the members of the Committee).

IBGE (2003), Norma de Serviço CDDI n.º 1, de 10.09.2003. (Regulation of the General Coordinator of the IBGE's Centre for Documentation and Dissemination of Information, n.1, of September 10, 2003 – that established the objectives of the rooms for use in the on-site restricted access).

Lei nº 5534, de 14 de novembro de 1968. Brasília, Diário Oficial da União. (Law 5534 of November 14, 1968. Law on the obligatory character of providing statistical data and confidentiality).

## ANNEX 1.16. CASE STUDY MICRODATA LABORATORY ANALYSIS – ITALY

### 1. Broad description

The Statistics Law 332/1989 allowed for the first time the Italian national statistical institute (ISTAT) to release microdata to external users. These are essentially data from social surveys where protection is chosen according to a statistical model; the complete list is on the Internet at [www.istat.it/servizi/infodati/index.html#standard](http://www.istat.it/servizi/infodati/index.html#standard). It was soon clear that this product was not able to cover all requests for access to microdata especially for research purposes. For this reason in 1999 ISTAT created the Laboratory for Analysis of Microdata (Laboratorio ADELE, an abbreviation of *Analisi Dati ELEMENTARI*), an on-site facility where researchers perform statistical analysis on confidential microdata files stemming from both social and business ISTAT surveys.

### 2. Why is it good practice?

Very often researchers need business microdata or social microdata with the maximum information content; therefore, restricting the detail in data (as in the released microdata file) is not a feasible solution. The restriction has to be made on the access, using administrative, legal, statistical and IT measures to avoid any breaches of confidentiality. In the Laboratory the users have access to the whole collection of validated microdata of the Institute with the maximum information content.

### 3. Target audience

Access to the Laboratory is allowed for research purposes only; projects are welcome from universities or research institutes or from bodies who can prove a recognized research attitude. Projects have also been accepted from ministries or national authorities who demonstrate that the proposed project has clear research intentions. Researchers from foreign universities and institutes are also allowed.

### 4. Detailed description

The Laboratory for Analysis of Microdata is located in Rome at the ISTAT premises; plans are in place to open new branches in regional offices of ISTAT in order to decentralise access.

A researcher or a team of researchers seeking access to one or more microdata sets must complete a form containing the following information: the institution where they work, the name of the person responsible for the research project (because, very often, research students themselves carry out the analysis of the data), description, rationale and objective of the study, data to be analysed, statistical methods chosen to analyse the data, statistical software needed, and the expected type of output to be taken away from the Laboratory.

All proposals are reviewed inside ISTAT to establish the admissibility and purpose of the request, the admissibility of the institution, the need for confidential data as opposed to available microdata products, the acceptability of the expected output with respect to confidentiality. This latter analysis is done to avoid fruitless analysis where it is known in advance that the type of output requested is far too detailed to be taken away from the Laboratory without a protection that will completely destroy the information content.

Data are provided in a safe setting: a room with PC on a network separated from the internal ISTAT one, where any input/output procedures are disabled to users. The most common statistical software is available and any other commercial statistical software brought by the user will be installed on production of a valid licence.

The users sign a contract that ties researchers to their institution and together they are responsible for maintaining confidentiality. In accordance with the Statistics Law, every research project is authorized by the President of ISTAT.

Access to the Laboratory is controlled and supervised and the final output of the research is released after checking for confidentiality by ISTAT staff. The results of the research cannot be considered official statistics.

The number of projects is increasing steadily every year; in 2005 ISTAT approved more than 30 projects with more than 100 days of work.

A complete description of the Laboratory together with the form to request access is available at [www.istat.it/dati/pubbsci/documenti/Documenti/doc\\_2004/2004\\_9.pdf](http://www.istat.it/dati/pubbsci/documenti/Documenti/doc_2004/2004_9.pdf)

## **5. Supporting legislation**

The renewal of the legal framework on privacy protection in Italy, finalised by the adoption of the Personal Data Protection Code (Law 196/2003), led to the development of the Code of Ethics and Good Conduct for any party that is in some way involved in the processing of personal information (journalists, historians, statisticians, police, health services, etc). The Code of Ethics and Good Conduct for Public Statistics (Provvedimento del Garante no. 13, 7/2002) has the status of law (Annex A.3 of the Personal Data Protection Code) and applies to all the processing of confidential data for the purpose of statistics and research in the framework of the National Statistical System. It sets criteria for assessing the identification risk of a statistical unit, the rules to be followed when providing information to subjects not involved in the National Statistical System (Laboratory and anonymous microdata - art. 7), and when exchanging confidential microdata inside the System, security measures and so on. The Statistics Law and the Code of Ethics provide a complete framework for access to microdata inside and outside the Statistical System.

## **6. Strengths**

Users can study the complete microdata set (except for direct identifiers) stemming from all social-demographic and business surveys as well as censuses conducted by ISTAT. There is no limit on the type of analysis that can be carried out at the Laboratory; this allows for more in-depth analysis of phenomena being studied, especially as far as business data are concerned.

## 7. Weakness

The Rome location of the Laboratory is a barrier for certain researchers; plans to add other laboratories in regional offices of ISTAT will only partly resolve this problem. The metadata and documentation are mostly provided in Italian and this represents a major problem for foreign researchers.

## 8. References

The Statistics Law (Legislative Decree no. 322, 6 September 1989), is available at [www.istat.it/dlgs322.pdf](http://www.istat.it/dlgs322.pdf)

Personal Data Protection Code (Legislative Decree no.196, 30 June 2003) available for download at [www.garanteprivacy.it/garante/document?ID=1169255](http://www.garanteprivacy.it/garante/document?ID=1169255)

*Code of Ethics and Good Conduct when processing personal data for the purposes of statistics and scientific research within the National Statistical System* also at [www.garanteprivacy.it/garante/document?ID=1169255](http://www.garanteprivacy.it/garante/document?ID=1169255), Annex A.3 pp. 130-141.

## ANNEX 1.17. CASE STUDY

### MANAGING DECISION MAKING ON CONFIDENTIALITY - SLOVENIA

#### 1. Broad Description

The Case Study presents the management issues at the Statistical Office of the Republic of Slovenia (SORS) associated with the release of microdata to researchers. Tasks of the Data Confidentiality Committee and the system of rules and procedures regarding the release of microdata to researchers are presented.

With the procedure described, the Director-General of SORS has the necessary advice before deciding on the release of microdata to researchers and all the researchers are treated in the equal standardised way.

#### 2. Why is it Good Practice?

It provides the necessary advice to the Director-General of SORS before deciding on the release of microdata.

Researchers are treated in equal way.

Trust in SORS regarding confidentiality of data is maintained.

SORS's staff is well informed about the procedure and can monitor the decision process and outcome which is applied in a routine fashion.

#### 3. Target Audience

Target audience are researchers in research and government bodies as well as individual researchers and social science data archives.

#### 4. Detailed Description

##### 4.1 Data Confidentiality Committee

The Data Confidentiality Committee deals with the problems of data confidentiality at SORS and was established as an advisory body of the Director-General.

The Committee has the following tasks:

- To take care of the implementation of the Rules on the Procedures and Measures for the Confidentiality of Data Collected Under the Program of Statistical Surveys by SORS
- To deal with various matters and to give advice to the Director-General of SORS regarding issues that cannot be solved by general rules from the field of data confidentiality
- To report to the Director-General of SORS and the Statistical Council of the Republic of Slovenia regarding the situation in the field of data protection at SORS.

With reference to its tasks, the Data Confidentiality Committee adopts findings, positions and opinions and forwards them to the Director-General of SORS. Members of the Data Confidentiality Committee are experts from SORS and authorised producers of national statistics as well as external data protection experts. Committee members are appointed by the Director-General of SORS.

## **4.2 The system of rules and procedures regarding the release of microdata to eligible users**

### **4.2.1 Organizational rules and procedures**

All requests received by SORS for the release of microdata are transmitted to the Data Confidentiality Committee, which prepares the opinion about the possibility of releasing the requested microdata and forwards it to the Director-General of SORS for approval. In preparing the data, subject matter specialists must take into account the »need to know« principle.

### **4.2.2 Rules for the release of non-anonymised microdata**

#### **a. Release of microdata within the system of national statistics**

Good practice: for implementing the Annual Program of Statistical Surveys, it is possible to exchange microdata between SORS and authorised producers of the program of statistical surveys.

#### **b. Release of microdata collected with combined questionnaires to partner institutions**

In some cases SORS sends together with one of the government institutions to the reporting units a combined questionnaire, thus decreasing the burden of reporting the same or similar microdata twice to government institutions.

Good practice: SORS collects microdata with a combined questionnaire together with a partner institution only if SORS and the partner institution have a legal basis to do this and if SORS's interest is not threatened by this method of microdata collection.

The legal basis for microdata collection by SORS and a partner institution must be printed on the questionnaire and the information letter must explain the purpose of microdata collection for the partner institutions.

#### **c. Release of microdata to observation units requesting own microdata**

In some cases observation units ask SORS for own microdata that they sent to SORS in the past.

Good practice: if SORS has these microdata, it transmits them to the observation unit within its technical and financial capacity. For the 2002 Population Census, SORS transmits prints of scanned census questionnaires.

- d. Release of microdata about their members to commercial and interest associations

To rationalise microdata collection and decrease the burden of reporting units, some commercial and interest associations do not collect microdata for various analyses themselves but ask SORS to transmit these microdata to them.

Good practice: SORS transmits individual microdata on a member of an association after obtaining written consent from the member.

- e. Release of microdata for the purpose of interviewing

For the purpose of interviewing, SORS may transmit to registered scientific research organizations and registered individual researchers only the following personal microdata: name and family name, address, year of birth, sex and occupation (see National Statistics Act).

#### **4.2.3 Rules for the release of anonymised microdata**

- a. Release of microdata to scientific research institutions and individual researchers

Good practice: microdata for scientific research and analytical purposes (secondary data analysis) are transmitted only to scientific research institutions and registered researchers that can prove their registration.

- b. Release of microdata to researchers in government bodies

Government bodies are statistical microdata users that have great and specific needs for microdata, so SORS facilitates their work regarding policymaking by enabling them to use microdata.

Good practice: microdata are transmitted to the government body if the purpose of microdata use is research or analysis.

The request is denied if the purpose of microdata use is to determine administrative advantages or disadvantages for business entities or natural persons.

- c. Release of microdata to social science data archives

By transmitting microdata to data archives, SORS enables analytical and research work.

Good practice: microdata transmitted to various social science data archives have the highest level of confidentiality based on the contract between SORS and the data archive.

## 5. Supporting Legislation

### 5.1 National Statistics Act

National statistics shall be implemented on the different principles among others on statistical confidentiality (Article 2).

The professional tasks performed by the Office within the framework of its basic functions shall among others develop methods and techniques for data protection (Article 7).

Dissemination of data shall be carried out in such a way that the persons or businesses cannot be identified (Article 34, 47).

For the purpose of conducting surveys, the Office may transmit to registered scientific research organizations and registered individual researchers only the following personal data: first name and family name of an individual, his/her place of residence, year of birth, sex and occupation (Article 41).

Statistics may be published in aggregate form only, by way of exception, data may also be published individually:

- upon written consent of the reporting unit as regards publication of the data in such a way;
- if data have been collected from public (generally accessible) data collections (records, registers, databases, etc.);
- if data are published in such a way that the person or business involved cannot be directly identified (Article 50).

### 5.2 Rules on procedures and measures for the protection of data collected through programmes of statistical research at the Statistical Office of the Republic of Slovenia

In communicating data to users, the principle of statistical confidentiality shall be respected. The principle of statistical confidentiality means that no data may be communicated to users outside the system of national statistics, which can be ascribed to a particular observation unit or which could indirectly enable this (Article 5)

Before communicating data referred to in the previous paragraph, the research organization or registered individual researcher shall sign the declaration on data protection (Article 16).

Data for research purposes may only be used by a registered research organization or registered individual researcher that has concluded an appropriate contract with the Office, which must contain the status of the user, the intended use of the data, protection of data and the manner and time of publication of the data (Article 17)

A proposal for concluding a contract shall be discussed by the Committee for Data Protection before the contract is concluded (Article 17).

## 6. Strengths

- It provides the necessary advice to the Director-General of SORS before deciding on the release of microdata.
- Researchers are treated in equal way.
- Rules and procedures for microdata release are transparent (they are published on intranet and internet)
- Procedures are easy to understand for the staff of SORS and researchers.
- Trust in SORS regarding confidentiality of data is maintained.
- Staff of SORS is adequately informed about the procedure for managing decision making on confidentiality and can monitor the decision process and outcome which is applied in a routine fashion.
- There are clear responsibilities for the upgrading of the system for managing decision making on confidentiality.

## 7. Weaknesses

- Time lag between the data request and approval

## 8. References

National Statistics Act ( [http://www.stat.si/doc/drzstat/ZAKON\\_O\\_DSTA\\_ENG.PDF](http://www.stat.si/doc/drzstat/ZAKON_O_DSTA_ENG.PDF) )

Description of access to microdata on SORS's home page

[http://www.stat.si/eng/drz\\_stat\\_mikro.asp](http://www.stat.si/eng/drz_stat_mikro.asp)

Rules on procedures and measures for the protection of data collected through programmes of statistical research at the Statistical Office of the Republic of Slovenia

[http://www.stat.si/doc/stat\\_urad/pravilniki/06-0471pravilnik\\_varstvo\\_podatkov\\_en.pdf](http://www.stat.si/doc/stat_urad/pravilniki/06-0471pravilnik_varstvo_podatkov_en.pdf)

## ANNEX 1.18. CASE STUDY

### MANAGING DECISION MAKING ON CONFIDENTIALITY - AUSTRALIA

#### 1. Broad description

Special governance arrangements have been put in place to ensure the Australian Statistician has sound advice before making decisions on whether to release anonymised microdata files.

The recommendation comes from the statistical area responsible for the statistical collection on which the microdata are based. They make judgements based on the demands from users of the information. It must be accompanied by:

- (a) a positive recommendation from a Microdata Review Panel (chaired by a senior methodologist) that the microdata are not identifiable; and
- (b) a positive recommendation from the Policy Secretariat area that the proposal conforms with policy and legislation requirements.

#### 2. Why is it good practice?

It provides the necessary assurances to the Australian Statistician before he makes decisions on release of microdata. In practice, he approves in principle the release of microdata from a particular collection (e.g. a Household Expenditure Survey). Australian legislation requires each release to an individual researcher to be approved. This responsibility has been delegated but only to senior executives.

#### 3. Target audience

Microdata releases are targeted to the research community, particularly those located in government agencies, universities and other research institutes.

#### 4. Detailed description

To assist uniformity in the process, standard templates have been developed for documentation.

The relevant statistical area will take the initiative in developing a proposal. For many surveys it will be a standard output, although there may still be consultation on the exact nature of the microdata release. In other cases, the decision to provide a microdata release could depend on representations from users and subsequent discussions with them.

The statistical area will then develop a proposal for a microdata release. It must first be cleared with a Microdata Review Panel. This is chaired by senior methodologists and they have developed criteria to assist them with their assessment. They will also do empirical analysis to help determine identifiability. If the result is unacceptable, they will make recommendations on how to reduce identifiability (e.g. by combining classifications). This will be done in collaboration with statistical areas.

Wherever possible consistent classifications on identifying variables such as age and occupation are used across different microdata releases. This simplifies the assessment task but is also of benefit to researchers working with several microdata sets.

The Microdata Review Panel is chaired by a senior methodologist. Its membership includes confidentiality experts and representatives from statistical areas.

One important component of the submission is the conditions that must be included in the Undertaking to be signed by the recipients of the microdata release. Some are prescribed in legislation; others can be determined by the Australian Bureau of Statistics (ABS) (e.g. non-matching with other data sets). The Microdata Review Panel may recommend conditions as part of their deliberations.

The next step is to get clearance from Policy Secretariat. They will check that the proposal conforms with legislation. They will also check that the proposal conforms with ABS policy on microdata release.

## **5. Supporting legislation**

The relevant legislation is described in Annex 1.1.

## **6. Strengths**

It provides appropriate checks and balances and the full range of information required for the Australian Statistician to make an informed decision.

## **7. Weaknesses**

The assessment of some proposals can be labour-intensive. Also the investigation may take effort. This can result in delays in decision making particularly when multiple proposals are being considered at the same time.

## ANNEX 1.19. CASE STUDY

### MICRODATA ACCESS IN THE OECD PROGRAMMES FOR INTERNATIONAL STUDENT ASSESSMENT (PISA)

#### 1. Broad Description

OECD's Programmes for International Student Assessment (PISA) are conducted every three years in order to collect student achievement indicators in a number of areas, such as reading, mathematics and science. Information is collected both on students and their schools. PISA is currently in its third project cycle: the first being conducted in 42 countries in 2000 (PISA 2000), the second in 41 countries in 2003 (PISA 2003), and the third cycle in 57 countries in 2006 (PISA2006). Data and instruments for all PISA cycles are available on the OECD PISA web site ([www.pisa.oecd.org](http://www.pisa.oecd.org)). Released data include microdata files, statistical tables published in the international reports, and special tables generated at the request of researchers.

#### 2. Why is it Good Practice?

An Agreement concerning the confidentiality in the use of PISA materials is established between the OECD and the countries participating in PISA. This agreement specifies that the use of all materials from the OECD/PISA is permitted solely for the national implementation of PISA in the participating country, preparation of national reports or documents, with the provision that no information derived from these materials shall be published or otherwise disseminated to any individual other than those identified in the Agreement prior to the publication of the first international PISA report by the OECD, or without prior consent from the OECD. The OECD reserves the right to terminate the Agreement at any time with immediate effect, for the reasons of failing to meet any requirements of the Agreement.

In the microdata files, student and school information are kept anonymous through the use of randomly assigned identification numbers and codes. This system has ensured anonymity while maintaining high levels of accuracy.

Released PISA data provide reliable and internationally comparable indicators that meet high technical standards. Only data concerning participating countries that have fully satisfied PISA Technical Standards in the areas of sampling (including population coverage, exclusions and response rates), translation and translation verification, test administration, quality monitoring, coding, data entry and data submission, are included in PISA international microdata files. Since a number of same items and instruments are used across all cycles, researchers have the opportunity to compare indicators over time.

The PISA microdata files are released publicly through the PISA web site in the first week of December in the year following that of data collection. This occurs concurrently with the release of the initial PISA international reports of each cycle. Researchers worldwide are immediately able to replicate the analysis presented in the international reports. Concerning PISA 2006, the data set is planned to be released on December 4, 2007. Participating countries are currently required to strictly maintain the embargo on releasing any results until that date.

### 3. Target Audience

All stakeholders involved in education: policy makers, researchers, teachers, school principals, parents and students.

### 4. Detailed Description

In the PISA web site, four data functions are available for each programme cycle:

- **Downloading of microdata files:** user can download questionnaires, code books, microdata files in TXT format, SPSS and SAS control files, and compendia. With these microdata files, researchers can conduct analysis and run models using statistical software such as SPSS and SAS. The PISA 2003 student microdata file includes more than 400 variables for approximately 276000 students. The PISA 2003 school microdata file includes 190 variables for approximately 10000 schools.
- **Interactive data selection:** user can construct tables by selecting countries and variables. Tables are immediately generated through the website. Included are estimates for the variable selected, student performance determined by the selected variable, as well as standard errors.
- **Multi-dimensional data request:** user can access more complex analytical results by selecting multiple variables. Results can be mailed directly to users through an email service based on the website.
- **PISA data service:** more advanced or customised analysis is available for a fee.

### 5. Strengths

All data files and data functions are available on the PISA website without requiring special registration. The entire data set is available publicly free of charge through the website. Various on-line functions (as described above) are available for handling and interpretation of PISA data. Users can select an on-line data function depending on their technical ability and the aim of analysis.

There is sufficient confidence in the arrangements for protecting confidentiality in the project. The high participation in PISA ensures the quality of the resulting statistics.

### 6. Weaknesses

Analysis of the PISA microdata may be complicated because it requires understanding and application of high-level statistical knowledge. In order to support researchers in conducting analysis, *PISA 2003 Data Analysis Manual* has been published. The Manual explains the methodological approach applied by PISA as well as SPSS and SAS macros and syntax for correct computation.

Since the data found in PISA microdata files can be quite extensive, computation time may be lengthy depending on a computer used. It is recommended that user defines cases selected and variables necessary for the analysis as narrowly as possible to ensure effective analysis.

## **7. References**

OECD (2005) PISA 2003 Data Analysis Manual. OECD. Paris.

Further information on PISA is available on the PISA web site ([www.pisa.oecd.org](http://www.pisa.oecd.org)).

## ANNEX 1.20. CASE STUDY

### POLICY ON INTERNATIONAL RELEASE OF MICRODATA - AUSTRALIA

#### 1. Broad description

The Australian Bureau of Statistics (ABS) is receiving a small, but growing, number of requests from overseas researchers for access to microdata.

There are two types of access sought: individual researchers seeking access for research projects, and requests to add Australian data to international databases. The Luxembourg Income Study (LIS) is a long-standing example of the latter type of request.

While granting of access to microdata to overseas researchers will remain a matter of judgement, a policy has been developed to provide guidance on when such requests might reasonably be considered.

#### 2. Why is it good practice?

Increasingly comparisons with other countries are being used to inform policy. It is why Australia is an active participant in organizations like OECD and other international collaborations with worthwhile objectives. Often these studies require access to microdata to achieve their research objectives.

The ABS can legally release microdata internationally under specified conditions. As well as legal requirements, there is the issue of public acceptability. To maintain the trust and confidence of respondents, there have to be assurances that their data is safe and being put to good use.

The policy statement provides a decision making framework to allow individual decisions on research access to be made on a consistent basis.

#### 3. Target audience

The international research community but particularly international agencies.

#### 4. Detailed description

For data to be released internationally, two key conditions should be fulfilled.

- (i) The study should be of interest to Australia. While this would always be a matter of judgement, some examples of work meeting this criteria might include: producing international comparisons in an area of topical interest; an overseas organization undertaking policy relevant work on behalf of Australia; methodological work that might lead to improved data collection practices and methods in Australia; and research that is relevant to Australian policy.
- (ii) The recipient organization and person should be trustworthy. While this also remains a matter of judgement a 'threshold' criteria may be that the organization has recognised international standing in the relevant field.

Unless the above two conditions are fulfilled, access should not be provided.

Even in these situations our Remote Access Data Laboratory (RADL) would be the preferred option if practicable. For requests made by individual researchers, access should only be granted through RADL.

The organization receiving the microdata may want to provide access to other researchers outside their organization to support the international study. They cannot do this legally. Each request should come to the ABS for consideration.

The process of granting access proceeds in two stages: an "approval in principle stage" which assesses the usefulness of the project and trustworthiness of the applicant, followed by an "approval" stage which involves signing of appropriate undertakings.

Where it is found that researchers or organizations have breached conditions of undertakings made, sanctions will be applied. Doing so can reduce the risk of further breaches by the relevant researchers/organization as it acts as a deterrent to others' breaching their undertakings.

There will be a graded series of sanctions as follows:

- 1) for minor breaches, issue a warning to the individual in breach of an undertaking and their organization (where there is suspicion rather than proof of a breach, this approach might be taken);
- 2) remove data access from the individual in breach of an undertaking, either in perpetuity or for a fixed length of time (e.g. three years);
- 3) remove data access from all researchers from the offending organization, or in the case of ABS microdata being part of an international study prohibiting further access to this data, either in perpetuity or for a fixed length of time (e.g. three years);
- 4) advising the researcher's managers, or other persons of authority, of the breach and the sanction;
- 5) publicising, to the relevant international research communities, that an organization has been in breach of their undertaking in relation to ABS microdata and that they are prohibited from using ABS microdata.

Which sanction to apply would remain a matter of judgement; however, the factors to consider would be:

- 1) whether the breach was intentional or not;
- 2) the nature of the breach;
- 3) the breadth of the breach (one researcher only as against multiple researchers);
- 4) the length of time that the breach had been occurring before detection.

## 5. Supporting legislation

Microdata is released under the provisions of Clause 7 of the Ministerial Determination (see References).

There is nothing in these provisions which prevents release to a person or organization residing outside Australia. We have been reluctant to do so because legal sanctions against breaches could not be applied.

Each release to a person or organization should be approved by the Australian Statistician or delegate within the ABS (at present the Deputy Australian Statisticians).

Although legal sanctions may not be possible, there are other sanctions that could be used. The most powerful (and easy to apply) is to withdraw access to all ABS microdata services.

## 6. Strengths

It provides a publicly defensible basis as to how Australia might participate in international research studies involving microdata. Previously, as an abundance of caution, we had provided virtually no microdata access to international researchers.

It provides a clear statement to the international research community on the ABS position.

It provides a clear statement of the ABS position to staff who might be collaborating with international researchers. They know what is allowable and what is not allowable and discussions can proceed on the basis of that understanding.

## 7. Weaknesses

The risk of actual identification is very small. The most likely breach is that the recipient of microdata may pass the microdata on to other researchers including those in Australia who are not authorised by the ABS to access the data.

If this happened, it might lead to perceptions about the security of microdata. This in turn could affect response rates and hence the quality of ABS statistical collections.

## 8. References

A copy of the policy statement can be obtained through [teresa.dickinson@abs.gov.au](mailto:teresa.dickinson@abs.gov.au).

## ANNEX 1.21. CASE STUDY MANAGEMENT OF RECORD LINKAGE PROJECTS – CANADA

### 1. Broad description

Since the mid-1980s, Statistics Canada has had in place a Record Linkage Policy designed to protect the privacy of individuals while, at the same time, permitting record linkage under certain circumstances. Record linkage can be undertaken for research and statistical purposes only, and where the public benefits of the proposed linkage are judged to outweigh the privacy intrusion inherent to the linkage. All record linkage proposals must follow a prescribed review process that culminates with approval by Policy Committee, the senior executive committee chaired by the Chief Statistician.

### 2. Why is it good practice?

The Policy ensures that an appropriate balance is maintained between two competing public goods: the public good resulting from information that can only be developed through record linkage; and the minimising of privacy intrusion – which, however, is inevitably involved at any time when information about people is used in ways that they have not authorised.

A standardized approach to record linkage is implemented throughout the agency. By following this strict protocol, Statistics Canada has avoided any negative public reaction that could jeopardize or interfere with the agency's current or future activities. Transparency, strong governing procedures and an ethical position on the undertaking of record linkages has led to the sound management of this important statistical activity, which can shed light on important issues of public interest, and has contributed to the maintenance of public trust in the agency.

### 3. Target audience

The Statistics Canada Record Linkage Policy applies to all proposed record linkage activities to be carried out by employees of Statistics Canada, regardless of the purpose or extent of the linkage activity.

### 4. Detailed description

The Record Linkage Policy provides a definition that captures all types of linkages. Record linkage is defined as the bringing together of two or more micro-records to form a composite record, where a micro-record contains information about an identifiable individual respondent or unit of observation, such as a person, family, household, dwelling, farm, company, business, establishment, institution, etc.

In deciding which applications to approve, Policy Committee looks for a high likelihood that the linkage would result in significant public benefits; a methodology that would yield valid results; and ensures that no disadvantage affecting the subjects of the linkage, individually or collectively, would result. In addition, for linkages thought to be especially sensitive, the Committee will seek out the view of the Privacy Commissioner(s), as well as the degree of public support from key client groups or other stakeholders. Furthermore, in order to ensure transparency, the Record Linkage Policy requires that all approved applications, and their expected public benefits, be listed on the agency's web site.

All record linkage proposals to Policy Committee must include the following information:

- A concise description of the intended linkage project and an outline of the proposed research plan. The purpose for undertaking the proposed record linkage must be fully discussed, including the key reasons for conducting the linkage and the intended use of the results. How the public interest is served must be clearly demonstrated, namely by asking and answering the question: “So what?” It is important to indicate whether the linkage study findings are to be used in the context of public policy development, adjustments to existing federal or provincial programmes (e.g. funding or administrative arrangements), administrative decision-making, programme or project evaluation, changes to medical procedures, improvements in workplace safety procedures and so on. Policy issues that may be supported by the results of the proposed linkage must also be identified. Where linkages involve the use of personal information, how the public interest benefits will outweigh any possible privacy intrusions must be demonstrated. Research projects that are dependent on the linkage must be described in detail, including the research hypotheses.
- An indication of whether the proposed linkage is once-only or ongoing.
- An indication of whether survey respondents or those involved in the study have provided consent for the record linkage activity, or have been notified of any intended record linkage activity. Direct approval by Policy Committee may not be required when informed consent has been obtained. The Director of the Division in Statistics Canada responsible for the implementation of the Record Linkage Policy has been mandated to determine whether fully informed consent was obtained, in which case the linkage project may proceed without further review, or whether special circumstances require that the linkage project be approved by Policy Committee. If obtaining consent is not feasible, any consultations or communication strategies with respondents, the target population, or the selected proxy representatives, if applicable, should be mentioned.
- An indication of whether a privacy impact assessment or an evaluation by an ethics review board has been conducted.
- An indication of any efficiencies or savings in terms of costs, resources, timeliness, and reduced response burden.
- The names, sources, and years of all the files to be linked are to be supplied. A summary of the file contents should also identify the variables from these files that will be used in the linkage.
- A detailed description of the methodology to be employed in the linkage, including a description of the models or statistical tests being undertaken, linkage techniques and any generalised linking systems that are to be used.
- In addressing the methodological issues, a discussion of the appropriateness of using record linkage as opposed to other methods. In this regard, it is especially important to highlight what other alternative sources were considered and why these were rejected in favour of record linkage.
- The ability of the data sources to support, with an appropriate level of statistical confidence, the expected findings of the research.

- Whether entire populations are being linked or whether only a sample is to be included. This is an especially important consideration as in some cases the privacy intrusion of a record linkage can be diminished by using a sample of the total population.
- Details regarding the outputs of the linked file.
- The maximum retention period for the composite file, after which the linkage file must be destroyed. In the event that a linkage project is not completed within the approved retention period, it is necessary to seek Policy Committee approval to retain the linked file for a longer period of time.
- In general, there is no analytical requirement to retain the identifiers on the linked composite file that is used for the data analysis. If there is a reason to retain the identifiers, an adequate justification must be provided.

Each submission must be accompanied by a one-page summary, which is included in Statistics Canada's *Annual Report to Parliament on Access to Information and Privacy*, and is also posted on the Statistics Canada web site.

## 5. Supporting legislation

The Record Linkage Policy is a major component of Statistics Canada's legislative and policy framework, and embodies several principles and provisions of the Statistics Act, the agency's governing legislation, as well as of the federal Privacy Act.

## 6. Strengths

The Policy ensures:

- that the trade-off between the expected public benefit and the degree of privacy invasion which may be involved is applied consistently across all linkages, projects and over time;
- that record linkages are carried out with great care, while pursuing selected key public interest objectives;
- that openness and transparency are maintained, from approval of the linkage to dissemination of the results;
- that every record linkage proposal is evaluated and approved based on its own merit, regardless of its source of funding;
- that for on-going linkages, the objectives are reassessed at set periods;
- that all analytic results are placed in the public domain and accessible to everyone;
- that linked files will be destroyed once the study is completed and the results released;
- that in the eventuality of a major public controversy, Statistics Canada would be in a position to convince Canadians that it had been very sensitive about their legitimate privacy concerns and gone to great lengths to minimize the intrusiveness of the linkage while still carrying out its mandate, thereby, hopefully, maintaining the public trust in the statistical office.

## 7. Weaknesses

- The Policy is viewed in some circles as being too conservative and in effect an impediment to research.
- The Policy sets out a rigorous review and approval process which involves the submission of documented proposals. Approval of record linkages may be seen as requiring an inordinate amount of time and effort.

## 8. References

Statistics Canada's Policy on Record Linkage is available on its web site at *Record linkage at Statistics Canada*. The web site also includes a summary of all approved record linkages, as well as a document on *Privacy-related policies and practices at Statistics Canada*, by Dr. Ivan Fellegi, Chief Statistician of Canada. See <http://www.statcan.ca/english/recrdlink/>

## ANNEX 1.22. CASE STUDY

### DATA LINKING WHEN PREPARING MICRODATA FOR RESEARCH - SWEDEN

#### 1. Broad description

When creating a statistical register for research, data linking is used for two different purposes:

- Different sources are combined to create an object set, the register population, with good coverage.
- Different sources are used to create the variables in the new register.

Different sources, such as administrative registers or pre-existing statistical registers, can consist of different object types. It may then be necessary to define the statistical units or objects in the new register in a suitable way so that data from sources with different kinds of units can be used together. Data linking can be used in the same way to combine microdata from sample surveys with data from administrative or statistical registers.

The case study is based on a report mentioned in section 8 below, which illustrates how different sources with agricultural data can be combined into a new kind of Farm Register with variables from many administrative and statistical registers. In the report data linking is discussed from a methodological point of view.

#### 2. Why is it good practice?

The microdata at National Statistical Offices have not been created to meet the needs of academic research. To meet these needs, new sets of microdata should be created, where existing sets of microdata are combined so that data sets with richer content are created. Exact linking with identifying variables is used to create this kind of microdata for research.

#### 3. Target audience

Persons at National Statistical Offices who prepare microdata for research and potential users of microdata for research.

#### 4. Detailed description

The original aim of the case study was to investigate how data linking can be used to create a new Farm Register at Statistics Sweden based on administrative data instead of a census. A new Farm Register can be linked to the Business Register in two steps:

- Integrating the census-based Farm Register and the administrative IACS Register with data from the Integrated Administrative and Control System, which is the system for agricultural subsidies used within the European Union.
- Linking the records in the integrated register with the records in the Business Register. After this linkage all variables in all statistical registers, which are linked to the Business Register, can be used to analyse the agricultural sector.

The role of the Business Register is to define the object set of all enterprises, including those belonging to the agricultural sector. To be able to create a set of microdata describing the agricultural sector, statistically interesting variables must be imported from other registers linked to the Business Register:

- Crop areas and subsidies of different kinds from the IACS Register. Persons employed by age and sex from the PAYE Register. The PAYE Register is based on the annual income verifications in which all employers provide information on wages paid to all persons employed.
- A large number of economic variables from the Register of Standardised Accounts, which is based on annual income statements from all firms: data from profit and loss statements, balance sheets, investments and labour costs.
- Turnover and other economic variables from the VAT Register, which is based on monthly or yearly VAT declarations from all firms.
- A large number of variables describing different kinds of vehicles owned by the agricultural unit from the Vehicle Register, which contains data about vehicles owned by businesses and individuals.

The conclusions of the case study can be summarised as follows:

The matching process must be carefully planned, considering which linking variables should be used, and in which order the different sources should be combined. The quality of the linking variables is important, and editing of these variables is an important part of the work. Causes and extent of mismatch should be investigated, and it must be decided if the non-matching units should be excluded or included in the register population. If they are included, mismatch will result in units with missing values for some variables. Seemingly matching objects should also be checked, since false hits will otherwise give rise to gross errors in data.

- The identifying variables should be edited before matching. Before editing of telephone numbers, only 47% of the farmers in the Farm Register could be matched to corresponding units in the IACS register. After corrections 64% could be matched.
- By combining two identifying variables (telephone number to the farm and the farm's tax identity number) the matching result is improved so that 96% of the units in the IACS register could be matched to units in the Farm Register.
- By combining these two identifying variables the matching result is improved so that more relevant agricultural units can be defined. Matching with only the farm's tax identity number resulted in (almost) only one-to-one matches between objects in IACS and the Farm Register. However, matching with both the tax identity number and telephone number resulted in a number of one-to-many matches and many-to-many matches. After data linking, new units should be created in the following way:
  - In some cases husband and wife, relatives or companions on the same holding make separate IACS applications for different parts of the holding's activities. As the relationships between these persons are informal and can change over the years, it is appropriate to combine all IACS applications and all legal units in the Business Register connected with these applications.
  - In other cases a number of holdings and IACS applications refer to the same telephone number. This is an indication that all objects have the same administration. If all holdings, all IACS applications and all legal units in the Business Register connected to the same group are combined, we will get an

agricultural unit, which can be described by all statistical variables in the register system.

- Linkages must be checked. First the one-to-one linkages were checked. A match between identification variables is not sufficient proof that the IACS and Farm Register objects are identical. If the IACS object has a larger crop area than the FR object this can indicate that the IACS object should be linked with two FR objects and vice versa. The linkages were checked by comparing total arable area, reliable crop area and location described by parish.
- It was found that there was serious under-coverage in the agricultural part of the Business Register. By combining the Business Register with the PAYE Register, the Register of Standardised Accounts and the VAT Register, this undercoverage was reduced from 25% to 3%. These administrative sources are not used by the Business Register today.

## 6. Strengths

By combining microdata from different sources, the relevance or scientific value of the data can be increased to a great extent.

## 7. Weaknesses

Mismatch and/or false hits will give rise to quality problems.

## 8. References

This case study is based on the following report:

Anders Wallgren and Britt Wallgren: Administrative Registers in an Efficient Statistical System – New Possibilities for Agricultural Statistics? How Can We Use Multiple Administrative Sources? Statistics Sweden and Eurostat 1999. The report is available at <http://www.scb.se/Grupp/Allmant/IACS2.pdf>

**ANNEX 1.23. CASE STUDY<sup>1</sup>****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](http://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.

---

<sup>1</sup> 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," Courier 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.



This should be read in conjunction with the **Glossary on Statistical Disclosure Control** developed by the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. (Available at: [www.unece.org/stats/documents/ece/ces/ge.46/2005/wp.45.e.pdf](http://www.unece.org/stats/documents/ece/ces/ge.46/2005/wp.45.e.pdf).)

### **National Statistical Office (NSO)**

Although the term is used in the singular, it is meant to incorporate all statistical agencies, or statistical units within government departments, who produce official statistics and provide access to microdata for statistical or research purposes.

### **Research community**

Although this mainly refers to people working in research institutions such as universities, it also includes researchers working in government agencies, NGOs, international agencies and the private sector. Some countries may want to define the research community more narrowly and only include those working in research institutions.

### **Statistical purposes**

It is particularly important to make a distinction between statistical and administrative uses. In the case of statistical use, individual data are used as an input to derive statistics that refer to a group of persons or legal entities. It may also incorporate support for other activities within a NSO (e.g. sample selection off a business register). Administrative uses concern decisions about a particular person or legal entity which may bring benefit or harm to the individual.

The statistics referred to above include statistical aggregates, statistical distributions, parameters for models and other forms of statistical analysis that may refer to groups of individuals or organizations without identifying them.

Microdata used for research is consistent with statistical purposes if it is being used to produce the type of statistics referred to in the previous paragraph.

### **Anonymised microdata files - Public Use Files**

These are microdata files that are disseminated for general public use. They have been anonymised and are often released on a medium such as CD-ROM sometimes through a data archive. The term anonymised implies that not only are names and addresses removed but that other steps are taken to ensure that identification of individuals is highly unlikely.

### **Anonymised microdata files - licensed files**

The term anonymised implies that not only are names and addresses removed but that other steps are taken to ensure that identification of individuals is highly unlikely.

Licensed files are distinct from Public Use Files in that use is restricted to approved researchers for approved purposes. A legal undertaking is signed before files are provided to them.

### **Remote Access Facilities**

These are facilities that provide researchers with the ability to produce statistical outputs from microdata through computer networks without researchers actually ‘seeing’ the microdata. The microdata itself does not leave the National Statistical Office. Remote Access Facilities may be of two types.

- (a) Remote execution where a researcher submits a programme and receives the output later by email.
- (b) Remote facilities where the researcher performs the analysis and can immediately see the answer on the screen.

### **Data laboratories**

This involves working on-site at the National Statistical Office, or one of its Branches, to obtain access to microdata. Access could be direct or indirect through staff of the National Statistical Offices. If access is direct, the researcher is in effect being treated as a temporary employee of the National Statistical Office with the inherent responsibilities.

### **Data swapping**

A disclosure control method for microdata that involves the swapping the values of records that match on selected records. The techniques maintain statistics such as means, variances and univariate distributions but can affect multivariate distributions.

### **Data perturbation**

Techniques for the release of microdata which change the data before dissemination in such a way that the disclosure risk for the microdata is decreased but the information content is retained as far as possible. Perturbation methods falsify the data by introducing an element of error purposely for confidentiality reasons. Possible perturbation methods are:

- rounding,
- addition of random noise.

### **Risk avoidance**

This approach tries to eliminate all risks. In the case of microdata confidentiality, it requires the confidentiality of the data to be absolute, not only in its own right, but in association with other available data.

### **Risk management**

Within the constraints provided by legislation, it involves identification of the risks and managing them in accordance with their significance (impact) and their likelihood. More effort is put into managing the high impact, strong likelihood risks. Microdata confidentiality may not be absolute when considered in association with other data. Confidentiality could be considered in association with other means of reducing the risk.

### **Data linking**

Data can be linked by exact matches (e.g. using an identifier such as name and address or ID number) or by statistical matches (using probabilistic matches). They may be NSO data sets only, a NSO and administrative data sets, or administrative data sets only. Data sets for a particular collection could be linked longitudinally. All these possibilities are incorporated within data linking.

\* \* \* \* \*