UNITED NATIONS **E**

Economic and Social Council

Distr.
General

ECE/CES/2006/6
30 March 2006

Original: ENGLISH

ECONOMIC COMMISSION FOR EUROPE **STATISTICAL COMMISSION**

CONFERENCE OF EUROPEAN STATISTICIANS

Fifty-fourth plenary session
Paris, 13-15 June 2006
Item 5 of the provisional agenda

GUIDELINES AND CORE PRINCIPLES FOR MANAGING STATISTICAL CONFIDENTIALITY AND MICRODATA ACCESS

Principles and guidelines of good practice – interim guidelines[1]

Submitted by the Task Force on Managing Confidentiality and Microdata Access

I. NOTE

1. These Guidelines have been prepared at the request of the Conference of European Statisticians (CES) by a Task Force set up for that purpose under the guidance of the Bureau of the CES. Dennis Trewin (the Australian Statistician) chaired the Task Force. The CES Bureau approved the Guidelines at its February 2006 meeting.

2. This document is referred to as Interim Guidelines as it has not been formally endorsed by the Conference of European Statisticians. It will be put forward for endorsement at the June 2006 meeting of the CES. It is also available on Internet at: http://www.unece.org/stats/documents/tfcm.htm.

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

[1] This paper has been prepared at the invitation of the secretariat.

Comments are welcome. To be considered by June 2006 CES, comments should be sent to tiina.luige@unece.org by 19 May 2006.

II. INTRODUCTION

4. 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 because of the 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 the academic researchers. International agencies are among those who want to use microdata for research purposes, particularly cross-country comparisons.

5. 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 that the 2003 Conference of European Statisticians (CES) agreed 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 that 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.)

6. There are two key objectives in these guidelines;
   (a) to foster greater uniformity of approach by countries whilst facilitating better access to microdata by the research community; and
   (b) through these guidelines and supporting case studies, to enable countries to improve their arrangements for providing access to microdata.

7. 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.

8. 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. 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.

9. 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.

10. 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 may be useful in determining appropriate changes. Indeed, in some countries, these guidelines may provide a useful stimulus for debating and agreeing on changes.

11. Any questions on these guidelines should be submitted by email to the Statistics Division of the UNECE at confidentiality@unece.org.

III. WHY SHOULD NATIONAL STATISTICAL OFFICES SUPPORT THE RESEARCH COMMUNITY?

12. 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 government's policies and programmes - they provide a mirror on society.

13. 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."

14. 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.

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

16. What is the research community? It includes those working in academic institutions of course. It also includes researchers working in non-government organisations 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 these Guidelines, the pertinent issues may vary somewhat between the different elements of the

research community.

17. The following sections try to bring together the perspectives of national statistical offices and the research community in the spirit of trying to find arrangements that largely satisfy the needs of both groups. This is considered in more detail in Chapter 6.

The perspective of the National Statistical Office

18. 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 the quality of official statistics.

19. 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.

20. 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.

21. NSOs may also be concerned about the 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.

22. 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 data are used more extensively in this way, it can provide an extra level of protection against budget reductions to these statistical programmes.

A. The perspective of the research community

23. 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:

(a) 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.

(b) 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.

(c) Broadly speaking, widely available access to microdata enables replication of important research.

(d) 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 to improve the quality of its surveys.

(e) It increases the range of outputs derived from statistical collections and hence the overall value for money of these collections.

24. 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. 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.

25. 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.

26. 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.

27. 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 we relied solely on improved statistical methods for confidentiality protection.

IV. CORE PRINCIPLES

28. 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.

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

(a) 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.

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

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

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

30. Making available microdata 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.

31. For Principle 2 a distinction has to be made between statistical or analytical purposes and administrative purposes and 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 systematically 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 discretion whether to provide access or not.

32. 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.

33. With respect to Principle 3, legal arrangements to protect confidentiality should be in place before any microdata can be 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 arrangements 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.

34. 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.

35. 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.

V. SUPPORTING LEGISLATION

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

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

37. 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.

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

(a) what can and cannot be done and for what purposes;
(b) the conditions of release; and
(c) the consequences if these conditions are breached.

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

VI. METHODS OF SUPPORTING THE RESEARCH COMMUNITY

40. 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 paragraphs. Case studies are used to further illustrate these different methods.

(a) 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) - Public Use Files (PUFs)	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 to ensure that identification of individuals is highly unlikely.) The level of confidentiality protection in Public Use Files, even when matched with other data files, should be such that an agreed undertaking is not necessary to ensure confidentiality. Public Use Files are a common way of providing access to researchers in many countries.
Anonymised Microdata Files - licensed files	Licensed files are also anonymised but distinct from Public Use Files in that use can be restricted to approved researchers and an undertaking or contract is signed before files are provided to them. 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, they may contain potentially identifiable data if linked with other data files; this is one reason why a preventive undertaking or contract is required.

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

Service	Notes
Remote Access Facilities (RAFs)	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.

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

Service	Notes
Data Laboratories (DL)	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.

A. Statistical tables and data cubes

41. 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.

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

43. 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.

B. Anonymised microdata files – Public Use Files

44. 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 AMFs and to put more reliance on facilities such as RAFs and data laboratories for researcher

access. The alternative is to put increased reliance on researchers honouring the undertakings or contracts that they make not to attempt to identify. Such undertakings are usually a key part of a release of licensed AMFs (see next section).

45. 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.

46. 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 undertaking may be one of the requirements of access. It should be possible to set up an arrangement where a prior undertaking 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.

47. 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.

48. 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, μ-ARGUS, is concerned with protection of microdata against disclosure. Several techniques are available in μ-ARGUS.

C. Anonymised microdata files - licensed files

49. This is an arrangement where specific users are authorised or licensed to use anonymised microdata files. There will be conditions associated with the licence, which can be specified in an 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. It may be desirable to have an undertaking or contract even when the availability of these microdata files is promoted for general use.

50. The conditions may include some or all of the following:
(a) an agreement by the researcher that he or she will abide by the conditions of release;
(b) that no attempt will be made to identify particular persons or organisations;
(c) the information will only be used for statistical or research purposes;
(d) the microdata will not be provided to other persons;
the microdata will be returned to the NSO when the research project is completed; and
(e) no attempt will be made to statistically match with other databases without

permission.

51. It is good practice for such an undertaking to have some legal standing, for example by being incorporated within enabling legislation. This would allow legal actions to be taken in respect of breaches. 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.

52. 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.

53. Some countries externally release microdata files containing confidential data albeit under strict licensing arrangements which specify the conditions under which the data can be used. This is not good practice unless it is covered by legislation. Even then there is a risk that if the public regards this use as inappropriate, it could lead to lower levels of cooperation in surveys.

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

D. Remote Access Facilities (RAFs)

55. These facilities are 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. Often there is a contractual arrangement between the NSO and the researcher or the institution of the researcher.

56. 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 developed. 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 put greater reliance on the agreements made by institutions and the retribution (particularly denial of future access) if there are breaches of the rules.

57. There are two basic types of RAF.

(a) Remote execution, where a researcher submits a program 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.

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.

58. 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.

59. 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.

60. 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 only to provide access to those researchers who have signed some form of agreement outlining the conditions of access. Education is also important, together with regular monitoring and checking of the use of these facilities.

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

E. Data laboratories

62. They 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.

63. Some NSOs 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.

64. 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.

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

F. Engaging a researcher as a temporary NSO staff member

66. Another way that researchers may access microdata is through engaging them as a temporary NSO staff member and making them subject to the same secrecy provisions as the staff of the NSO. This should really be seen as supporting the NSO's work programme rather than as a data access solution. It should not be done unless the researcher is assisting with the work of the NSO, otherwise it could be seen as a sham. If this type of pretence were occurring and became public, confidence in the NSO would diminish.

67. 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. But the NSO must accept the merit of the proposal and incorporation in its work programme activities. 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).

G. Business data

68. 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. In some countries, databases of business data are often 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 them 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.

69. 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.

70. In terms of the dissemination streams:

(a) Statistical tables remain relevant, although the higher level of identification risk means that more detailed data will generally not be available for businesses.

(b) Anonymised microdata files may only be relevant for the smallest businesses. For some research, this 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. Anonymised microdata files are likely to be of limited use.

(c) 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.

(d) Data laboratory arrangements are likely to be most pertinent for access to microdata files of businesses. Such arrangements exist in Statistics Netherlands, for example.

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

VII. MANAGING TENSIONS BETWEEN NATIONAL STATISTICAL OFFICES AND RESEARCHERS

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

72. 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.

73. 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 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. (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 more unique cases.)

74. 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 a public debate. But general community concerns about privacy suggests 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.

75. Transparency is important 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 certain 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 fundamentalists could sway public opinion. Support from respected and authoritative persons is very important.

B. How do NSOs manage the risks?

76. Some suggestions are outlined below:

(a) Agree on a set of principles that should be followed in the provision of access to microdata (such as those outlined in Chapter 3).

(b) Ensure there is a sound legal and ethical base (as well as the technical and methodological tools) for protecting confidentiality with microdata access. This legal and ethical base requires a balanced assessment between the public goods 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 the research proposal and the credibility of the researcher, and there should be some allowance for this in the ethical base. Access should not be regarded as automatic. NSOs must conform with the legislation or other protocols that operate in their country, regardless of the merit of the proposal.

(c) Have an arms-length process for the balancing of 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. Ethics committees may also 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.

(d) Be completely transparent about specific uses of microdata to avoid suspicions of misuse.

(e) Be prepared to provide more access through remote access facilities and data laboratories as 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.

(f) 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.

77. 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 of 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. 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.

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

(a) Asking them to prove their bona fides as researchers and to demonstrate the public benefits of their research and that the microdata are 'fit for purpose'.

(b) Making them sign a legally binding undertaking with similar penalties to those operating for NSO staff if they breach confidentiality provisions.

(c) 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.

(d) 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). Undertaking legal action where appropriate.

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 and they still must be followed.

C. Other issues

80. It is sometimes argued that respondent consent should be sought before release of microdata outside of the NSO. This should be discouraged, as:

(a) There are significant practical issues associated with seeking and managing consent.

(b) Data being provided are unidentifiable and are only being used for statistical purposes, consistent with the purpose of the data collection.

(c) 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. By this means, it can be argued that passive consent has been obtained.

(Note: 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.)

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.

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 defenses?

(a) 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.

(b) 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.

(c) NSOs should point to the overall public benefits of providing microdata access, particularly for the situation where the offence has occurred.

(d) 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.

VIII. MANAGEMENT ISSUES ASSOCIATED WITH THE RELEASE OF MICRODATA

A. 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:

(a) the risk of identification is sufficiently small;

(b) the adjustments made to the data items have not unduly damaged the microdata file for research purposes; and

(c) 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 16 and 17 describe the arrangements in Slovenia and Australia respectively.

B. Managing metadata

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

(a) a description of the survey including any information on quality;

(b) a list of the data items and the classifications used (sometimes referred to as a ‘data dictionary’); and

(c) definitions of the data items.

Provision of (a) 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.

C. 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 are more likely to occur 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 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:

(a) the institution has taken appropriate steps in dealing with the offence committed by the researcher; and

(b) 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.

IX. SOME SPECIAL ISSUES

A. 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 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 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:

(a) 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:
(b) Public Use Files where they exist;
(c) Licensed Anonymised Microdata Files, where countries are able to do this;
(d) Remote Access Facilities with appropriate safeguards;
(e) Collaboration with a researcher based in the NSO or the NSO's country.

100. From the point of view of microdata access, surveys of type (a) 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. 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 (b), Public Use Files are only available for some countries.

Licensed anonymised microdata files (i.e. approach (c)) 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:

(a) trust in the researcher and his or her institution;
(b) whether the study is of importance to the country; and
(c) 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 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 to release to researchers via the NSO of the researcher's home country. But in all cases, 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:

(a) limiting access to particular divisions of the international agencies and prohibiting the passing on of data to others;
(b) limiting the purposes for which the microdata could be used without permission;
(c) that the microdata must be returned on request (e.g. it may contain errors);
(d) the opportunity to comment on published material which uses the microdata; and
(e) 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 (d)) 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)). 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. 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 18 describes the arrangements used in the OECD PISA study.

B. 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 19 and 20 describe how Canada and Sweden respectively manage data-linking arrangements in their countries.

X. LIST OF CASE STUDIES:

* * * * *