STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CES/BUR.2005/11
1 October 2005

CONFERENCE OF EUROPEAN STATISTICIANS

First meeting of the 2004/2005 Bureau
Washington, D.C., 18-19 October 2004

Item 10: Report of the Task Force on Statistical Confidentiality and Microdata

CES TASK FORCE ON CONFIDENTIALITY AND MICRODATA

Note prepared by Dennis Trewin, Chairman, Task Force on Confidentiality and Microdata

BACKGROUND

1. The Task Force prepared a Discussion Paper. It was one of the papers considered at the 2004 Conference of European Statisticians (CES WP.1). It has been distributed further and a number of useful comments were obtained. The main purpose of this paper is to summarise those comments and to outline some specific questions for the consideration of the CES Bureau. I have not obtained the comments from fellow Task Force members in the preparation of this paper. However, I will seek their comments over the next few weeks and inject them into the CES Bureau discussion.

2. The summary is in two parts. First, some specific questions were asked in the Discussion Paper and the responses to those questions are summarised in Section 2. In Section 3, I have summarised the other main points made by commentators. In Section 4, I have summarised the main areas of agreement. In Section 5, I have outlined the areas of disagreement and some questions for the CES Bureau to consider. Finally, in Section 6, I propose "next steps" with the objective of preparing a final document for CES Task Force.

COMMENTS ON THE QUESTIONS POSED IN THE DISCUSSION PAPER

3. The overall tone of comments was positive about the directions taken in the paper. Comments were at the margin rather than fundamentally disagreeing with the proposals in the paper. The questions posed in the Discussion Paper are shown in italics below.

4. *The majority of NSOs have taken a very cautious approach on confidentiality to the extent of virtually avoiding all risks. Developments in technology, and the increasing availability of public and private data bases on individuals, suggest we take an even more cautious approach to avoid the release of unidentifiable data. However, the general feeling of the 2003 CES suggested that this may not be the sensible approach to take if you balance these two public goods. As NSOs should not go beyond their legal and other obligations, this may require countries to change their legal and other arrangements for providing access to microdata in order to provide a more appropriate balance. Is it accepted that, in making arrangements for access to microdata, these two public goods need to be balanced?*

5. There was a mixture of views, even within the one Office. Generally, it was agreed that there should be balancing between the two public goods but never in a way that identifiable data was available to researchers. No-one argued against researcher access to (unidentified) microdata for research purposes. In fact, the sentiment among many NSOs is that they should try to increase access, compared with the past, but in ways that protect confidentiality of data provided to NSOs. The form of balancing

the two public goods may vary somewhat from country to country but whatever is agreed should be enshrined in legislation, complimented by processes which are documented and transparent.

6. *Section 6 of the paper discusses how the tensions between the NSO and Researcher perspectives might be resolved. Is this supported?*

7. In para 27 of the Discussion Paper, some suggestions were made on how NSOs should manage the risks. These are repeated in the next paragraph and were largely accepted. Any proposed changes are shown in square brackets. The issue of greatest concern is where there is a need to assess the merits of specific proposals. Some of the commentators are concerned about their ability to make these judgements on an informed basis. Nevertheless, there was general agreement that NSOs should adopt a risk management rather than a risk avoidance approach.

8. How do NSOs manage the risks?

i) Agree on a set of principles which must be followed in the provision of access to microdata.

ii) Ensure there is a sound legal and ethical base (as well as the technical and methodological tools) for protecting confidentiality through microdata access. This legal and ethical base requires a balanced assessment between the public goods of confidentiality protection on the one hand, and public benefits from research on the other. This will depend to a large extent on the merits of the research proposal and the credibility of the researcher. [Not all were comfortable with their ability to assess the merits of research proposals.]

iii) To have an arms length process for the balancing of the public good which might be derived from access to confidential data versus the risk of confidentiality violation. Ethics Committees may be able to assist in situations where there is discretion in deciding whether to provide access or not but regardless NSOs must conform with the legislation or other protocols that operate in their country. [The support of Privacy Commissions or equivalent bodies to the process was important.]

iv) Be completely transparent about the specific uses of microdata to avoid suspicions of misuse. [Of course, this is not possible for Public Use Files.]

v) Provide more access through remote access facilities and data laboratories as completely [because of statistical matching possibilities] unidentifiable microdata for public release may no longer be possible without considerable "distortion" of the data. Explore other opportunities to use technologies to improve access to microdata in ways that adequate confidentiality protection is provided.

vi) Put some of the onus of responsibility to the research community. Ensure researchers are aware of the consequences to them and their institution if there are breaches. Follow through on retribution if there are breaches. Access should be regarded as a privilege not a right. [Encourage the development of a Code of Conduct for the academic research community, endorsed by universities and other research institutes. Report breaches to the relevant bodies including Research Councils. The damage to a researcher's reputation through such steps should be a sufficient disincentive.]

9. Some countries with less developed statistical systems believe the risks are too great at this time of their development - researchers will be under pressure to pass on otherwise confidential data.

10. The Discussion Paper also posed the question of how NSOs can put some of the risk back on to researchers. The concept was generally accepted. Any proposed changes to the Discussion Paper are shown in square brackets.

i) Asking them to prove their bona fides as a researcher. Demonstrating the public benefits of their research. [Some NSOs have concerns about their ability to judge the merits of different proposals.]

ii) Signing a legally binding undertaking with similar penalties to those operating for NSO staff if they breach confidentiality provisions.

iii) Ensuring researchers are fully aware of their obligations through appropriate training. Follow-up with effective audit and monitoring procedures.

iv) Where offences occur, withdrawing all current and future services from the researcher and possibly their institution for a period of time (possibly until they have undertaken disciplinary action against the offender). Undertaking legal action where appropriate. [Make legal action a fifth and distinctive point.]

11. *Section 11 suggests that the access arrangements should be different for data about businesses but the underlying principles remain the same. Is that supported?*

12. Confidentiality protection is just as important for business data as personal data, although the nature of concerns of businesses (eg market sensitivity) are different to those for persons. It was also pointed out that for many countries there is some information already in the public domain about businesses that could possibly be used for statistical research purposes.

13. Overall the solution was seen to be in Data Laboratory arrangements. Several countries have successfully introduced such arrangements for business data. These can be expensive to operate so there is a need to be more selective in which research proposals to support. Remote access arrangements are more difficult with microdata about businesses than microdata about persons because the confidentialisation procedures are more complex. But it may be possible to provide access to microdata files that are limited to small businesses.

14. *Some draft principles are outlined in Section 13. Are they supported? Should they be extended?*

15. The Principles as proposed in the Discussion Paper are as follows.

Principle 1: It is appropriate for microdata collected for official statistical purposes to be used for secondary data analysis to support research as long as there are prescribed conditions that protect confidentiality.

Principle 2: There should be a legal or other arrangement to support use of microdata in order to increase public confidence that microdata will be used appropriately. Provision of microdata should then be consistent with these legal and other arrangements.

Principle 3: Microdata should only be made available for research or statistical purposes.

Principle 4: The processes for researcher access to microdata as well as the uses and users of microdata (except for public use files) should be transparent, and publicly available, again to increase public confidence that microdata is being used appropriately.

16. For Principle 1, some commentators felt these went further than implied by the Principle 6 of the United Nations Fundamental Principles of Official Statistics which states that "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". I propose to substitute "statistical" for "secondary data" but it begs the question of what is meant by statistical analysis. This will need to be spelt out in the Guidelines because clearly commentators are making different interpretations.

17. For similar reasons, "research" should be removed from Principle 3. It was pointed out that there is nothing to stop Public Use Files being used for non-statistical purposes (except the limitations within the data itself).

18. Principles 2 and 4 were supported with recognition that the way they are implemented would vary from country to country. It was suggested that the Principles should include a ban on redissemination (ie researchers gaining access through other researchers).

19. *Section 15 outlines some preliminary thoughts on access by international researchers. Any comments?*

20. There was general agreement that it was important to support international research, particularly to support comparisons across countries. Although the risks were low, the general agreement was that anonymised microdata files (as distinct from public use files) should only be released with some legal protection. Some suggested the establishment of an international law. I am not sure how practical that would be. For EU countries one already exists. A more realistic alternative is to release through the NSO of the host country of the researcher and use the legal protections provided by their law. Most countries are comfortable with this (and some are already doing so) but not all countries have legislation in place that enables microdata release with adequate protection.

21. Remote access facilities are an alternative way of providing international researcher or international agency access to microdata. International agencies, who are not happy with remote access arrangements as the means of accessing microdata, would need to come to some arrangement with the country in which they reside, providing sufficient countries are prepared to provide their microdata in this way.

22. This topic was discussed in June 2004 at the first meeting of the OECD Committee on Statistics. Their conclusions were:

"In the discussion, many delegates underlined the usefulness of examining micro-level data for research and analysis and some countries pointed to recent efforts to make such data available for research at the national level. At the same time, there was a nearly unanimous view that bringing together microdata at the international level posed difficult issues of data confidentiality and data treatment. Rendering micro-level data fit for analysis and maintaining it on an ongoing basis was judged to be expensive and therefore, despite its usefulness, it was not clear whether benefits exceeded costs. Thus, at the present stage, the project only received very moderate support. However, the Secretariat could explore further the idea, and might present a more precise proposal for a possible future international Microdata Centre."

23. *Are the proposed principles appropriate for linked data sets? Should any additional principles apply?*

24. It was agreed that the same Principles should apply but noting that the risks would be considerably higher with linked data sets. Public perceptions also have to be addressed. A particular risk is if the owner of one of the linked data set was able to get access to the linked file. Identification of the information would be straightforward in this case. For this and other reasons some countries would not let linked data files leave their organisation. Other means of access would have to be developed such as remote access facilities.

25. Linked data files are a potentially powerful way of providing new forms of official statistics. Yet in most countries they have been rarely used if at all. It is a topic that is well worth further discussion and debate within the CES Forum, perhaps as a subsequent stage of the work of this Task Force. Statistics Canada has a longstanding policy which might be a useful starting point for discussion.

26. *Should the provision of Public Use Files be discouraged unless there is some form of undertaking by the person accessing the file?*

27. In those countries where they are available, researchers see these as being of great value. Also, from the point of view of NSOs, administrative costs are much less. However, there was a wide variety of views among countries. Some were very supportive (eg USA) and did not see the need for an undertaking whereas other countries (eg Australia) were quite negative because of the identification risk.

28. The main identification risk lies through doing a statistical match between the public use data file and other files accessible to the researcher. If data is available about household structure, and the demographics of other household members, research done in Australia suggests that the identification risk may be much greater than perceived even though you are only applying statistical matches.

29. The identification risk will be reduced through:

- The size of population of the country - it is much less for larger countries;
- The availability of other data bases containing individual household data;
- The amount of geographic detail released;
- The amount of demographic detail released;
- Whether an undertaking (not to identify) is made by the researcher;
- The age of the data on the Public Use File - the identification risk will be less with older data set so one strategy might be to only make older microdata sets available as Public Use Files.

30. Ultimately, it is a decision for each country cognisant of the risks that exist within their own particular circumstances. But for smaller countries, the risks of identification could be quite considerable.

31. *Apart from Public Use Files should equality of access be a principle for the provision of microdata? Alternatively, should discretion be provided to the NSO?*

32. There were a variety of views. Some thought equality of access was a basic principle of official statistics and this also should apply to access to microdata. But the majority thought the NSO should have some discretion when it came to release of microdata - there may be users and uses for which it would be inappropriate to provide access to microdata. However, they thought the basis for using discretion should be formalised in transparent guidelines.

33. *Should the use of microdata files for non-statistical purposes be banned?*
34. There is agreement that microdata files should not be used for non-statistical purposes. This would be contrary to the Fundamental Principle of Official Statistics? In addition, it would be inconsistent with the compact we made with respondents at the time of data collection. The exception would be when informed consent was obtained - this should only be sought where there is a clear benefit to the respondent.

35. As mentioned before, an issue of debate is what does "statistical purposes" mean? We should make some explanation in the final document although it may still be open to some interpretation. Statistical purposes could include a range of statistical outputs such as aggregate tables, models and visual representations (that don't identify individual persons or organisations).

36. *If public good is the main reason for providing microdata services how important is it that research based on microdata files be put in the public domain? Public use files would be an exception.*

37. There was general agreement that this should be a condition if publicly used resources were going to be used in this way. However, it often should not be the NSO who is the publisher. Often, the nature of findings will be such that the NSO should not be directly connected, or they may have concerns about quality. Nevertheless, NSOs could facilitate putting this information in the public domain by hosting Conferences of researchers or similar events.

OTHER SIGNIFICANT POINTS MADE BY COMMENTATORS

38. A number of other points were made. they are summarised briefly below. Some of the more important points are picked up in the next two Sections.

(i) The precise arrangements will vary by country - among other things they will depend on their culture, legislation and overall capability.

(ii) NSOs should be funded to support this research work - either through direct budget appropriations or through researchers providing compensation for the NSO costs in supporting their needs.

(iii) Should we treat all researchers the same? For example, some suggested that researchers in universities should be given access to microdata whereas researchers in government agencies. Others had a different view pointing out that researchers in government agencies are more likely to uphold undertakings.

(iv) Although most countries agreed with a risk management approach, some were concerned about the loss of control of "contracting out" confidentiality protection.

(v) There was general agreement that remote access facilities will play an increasingly important role.

(vi) Development work should continue on improved ways of protecting the confidentiality of data.

(vii) It is not just the quality of data that is important but the quality of documentation including meta data.

(viii) The paper could do more to acknowledge the damage that a single incident can cause.

39. It has been pointed out that there was not much from a respondent perspective in the Discussion Paper. Some work has been done in Australia to obtain that perspective using focus group views on a potential Data Linkage project. A number of important points have emerged from these discussions.

(i) Respondents support use of their microdata for worthwhile purposes. However, they would like some assessment of the merits of research projects.

(ii) They also want the uses to be transparent not secret.

(iii) A small but significant proportion were concerned about linking of ABS and external administrative data files (because they perceived there was a risk of information flowing in two directions). They were generally supportive of the linking of files already held by the ABS.

40. This is generally consistent with the arguments of the Australian Privacy Commission on data linking projects. For them, to support linking projects, there has to be clear public benefits, supporting legislation, a secure technical environment and transparency.

ISSUES ON WHICH THERE WAS BROAD AGREEMENT

41. The areas of broad agreement are set out below. Of course there were also some areas of disagreement. These are outlined in the following.

(i) It is appropriate for NSOs to provide access to microdata for statistical research purposes. But NSOs should have discretion over whether to provide this access or not.

(ii) A risk management rather than a risk avoidance approach should be adopted. Although NSOs bear much of the responsibility for managing the risks, some of it should be shared with researchers. The paper describes ways in which the responsibility can be shared. There must be real "punishment" for those researchers who breach the conditions under which they were provided access.

(iii) There is a need for transparent guidelines for researcher access to microdata - the principles in the paper (modified slightly) provide a starting point for establishing such guidelines. They should be supported by enabling legislation which also outlines the conditions and constraints.

(iv) The guidelines will vary somewhat from country to country. Among other things it will depend on the legislation in place, the culture that exists with respect to privacy and confidentiality, and the maturity of the statistical system.

(v) Linked data files have to be handled with much greater sensitivity.

(vi) There is a need for strong supporting infrastructure (metadata, researcher training, etc) when microdata is made available to researchers.

(vii) There was general agreement that increased use of data laboratories and remote access facilities are the way of the future and should be encouraged. This was particularly the case with business data where it was generally not possible to provide an anonymised data file. (The ABS is looking at developing an open source version of its remote access facilities.)

(viii) The need to provide international access to microdata is agreed - some comparisons across countries do depend on researcher access or international agency to microdata. However, there are a variety of views on the best means to do this (discussed below).

(ix) It was generally agreed that researchers should put their findings in the public domain as they are using a publicly funded resource. Although the contribution of the NSO should be recognised, it was felt that the NSO should not be the publisher in most cases. There may be tensions with the NSOs need to be seen as politically independent. Ensuring quality was another concern.

ISSUES ON WHICH A RANGE OF OPINIONS WERE EXPRESSED

42. It is suggested that the CES Bureau should specifically discuss these points as there is some difference of opinion.

(i) Whether data should be used for statistical purposes only. There were a variety of views expressed some of which may be due to different interpretation of the term "statistical purposes". The different views are:

(a) microdata should not be used for research purposes, just for producing statistics;

(b) microdata could be used for research purposes using statistical models, analysis and data based on the microdata;

(c) microdata should be available for broader purposes than research.

Firstly the paper should clarify the meaning of the terms "statistical purposes" and "research". Given that, my view is that (b) above is appropriate. The exception of course is Public Use Files where, by their nature, there is no control on how they will be used.

(ii) There were a variety of views of whether all types of researchers should have the same access - many felt that government researchers should hot have the same access as academic researchers as the risks may be different. There are others who felt that government researchers are more likely to conform with their undertakings and obligations.

This is a key point of difference and may not be able to be resolved. It depends on the culture that exists in the different countries.

(iii) Another issue on which there was some confusion is "informed consent". Most interpret this as meaning each individual respondent gives consent. Others interpret this as meaning all respondents are given information that advises them that their data may be used for research purposes.

The informed consent of each respondent is usually impractical. But NSOs should be transparent about the potential use of microdata for research purposes.

(iv) Should there be equality of access? A possible different treatment of researchers, depending on their institutional arrangements, is mentioned in (ii). There is also the issue of whether the proposal should be based on the potential public benefits of the proposal and/or the merits of the researcher.

In my view the NSO should have discretion. I think it should be based on the potential public benefits. Ethics Committees (or similar bodies) may be able to assist the NSOs in making these decisions. I would avoid making judgements on the merits of individual researchers unless ethnical concerns about the researcher have arisen in the past.

(v) Should there be Public Use Files? Should there be a signed undertaking made as part of the access arrangements for Public Use Files?

This is a choice for individual countries but they should be cognisant of the risks. Paragraph 28 describes how they can be minimised. Generally, the risks will be considerably greater for smaller countries.

(vi) The need for access by international researchers and international agencies is understood and supported. But there is no agreement on how to provide this access.

I think there are four ways, probably in order of convenience to the researchers. First, if it is an international agency, they could prescribe a law to protect confidentiality. Eurostat are in this position but this is not the case for many other international agencies. Second, a national statistical office (with adequate legal protection) could be the data host and undertake the international research. This type of approach has been used with the Adult Literacy Surveys. Third, microdata files could be provided through the NSO in the country of the researcher, using legal protection provided by the laws of that NSO. Fourth, remote access arrangements could be used. In the case of international agencies, even if remote access arrangements are used, there would still be merit in a centre to administer the arrangements, maintain the metadata, etc.

(vii) Although this was not specifically raised in the Discussion Paper, an issue worthy of debate is whether there should be provision to allow researchers as a deemed employee even though they were

not doing work for the National Statistical Office. Deemed employees would still be subject to the same confidentiality provisions and penalties as other staff of the NSO.

These arrangements could be supported if the work was clearly statistical in nature, had strong public benefits and the arrangements were transparent.

NEXT STEPS

43. The next step is to produce the final document. It has been commented that the Principles by themselves do not provide sufficient guidance. Therefore the final product of our work should be in the way of Guidelines on the Provision of Microdata Access for Research Purposes but it would include the Principles. It will also make reference to how to manage access to business data and linked data. It would also include some case studies of good practice.

44. It is hoped to have an advanced draft ready for discussion by the CES Bureau next February. This will have incorporated the comments of the Task Force members. The case studies will not be complete but a list of proposed case studies will be provided. In the first instance, CES Task Force members will be asked to identify suitable case studies.

45. Although we are only looking at Guidelines that have the imprimatur of the CES, the UN Statistical Division is interested in incorporating more broadly.

ITEMS FOR DISCUSSION

46. Whilst the Bureau may want to discuss any of the issues raised in this paper, it is suggested that Sections 4, 5 and 6 should provide the focus of discussion.

* * * * *