UNITED NATIONS

E

Economic and Social Council

Distr.
GENERAL

CES/2003/12
7 April 2003

ORIGINAL: ENGLISH

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Fifty-first plenary session
(Geneva, 10-12 June 2003)

ACCESS TO MICRODATA – ISSUES, ORGANISATION AND APPROACHES

Paper submitted by Australian Bureau of Statistics (ABS)[1]

I. BACKGROUND

1. Ensuring confidentiality is not only important for legal and ethical reasons, but to maintain public trust. The increasing demand for detailed data, combined with the increasing power and capability of technology, and the availability of potentially matchable data sets, makes the challenge of maintaining the confidentiality of microdata more difficult. It is virtually impossible to release microdata which contains useful information that is unlikely to be unidentifiable. Longitudinal data sets increase the problem. We can no longer rely solely on different forms of data amendments to manage disclosure risks.

2. On the other hand, there is increasing demand for access to microdata to support a range of research and secondary data analysis. Increased computing power increases the capability of researchers to undertake these types of analysis.

3. There are several motivations for addressing the issue on how to best provide researcher access to microdata.

- Valuable (and high quality) data is underutilised.
- Researchers may try to collect substitute data sets in order to obtain microdata, which is a waste of public resources (to obtain what is probably lower quality data).

[1] Prepared by Dennis Trewin.

- Government agencies may look to use alternative data providers to obtain survey data for research and analysis purposes, resulting in lower quality data that may not contribute to national statistics.

4. There is another important element that we need to consider - the incredible potentially valuable analytical power of linked data sets; including links with ABS data sets.

5. This range of factors has led us to rethink how we provide access to microdata. This is true for many other NSOs, many of whom are in the process of changing their practices. The steps we have taken, or plan to take, are described in this paper. Different strategies may be required for household and business based surveys. This paper only attempts to describe the ABS situation but hopefully this will be relevant to the situation many other NSOs face.

6. Before moving on, it is worth emphasising that whatever is done must be both legal and publicly acceptable. The law could be changed but this is not a quick or straightforward process and may raise unnecessary concerns. Consequently our approach is to work within existing law.

II. A BRIEF DESCRIPTION OF DEMAND

7. Ideally users would like:

- the ability to work interactively with the data;
- access to ABS experts and good documentation to describe the data;
- an increasing number of data sets available;
- good quality data for populations and variables of interest and some information about the sources of error;
- timely releases; and
- increasingly, access to linked data set (including data linked over time).

8. There has been a consistent message from researchers that we have taken too conservative an approach to the release of microdata. As a result, a recent focus has been to consider how we can increase access to microdata, while maintaining our high reputation for safeguarding privacy (and staying within the law) which is so important for maintaining a high level of cooperation in our surveys.

9. The use of linked data sets raises the possibility of ABS acting as a custodian of non-ABS data sets to ensure that there is appropriate confidentiality protection. While this is entirely consistent with National Statistical Service objectives, appropriate policies and operational procedures need to be developed. This is discussed further in paragraphs 24-28.

III. MEANS OF SATISFYING THE DEMAND

10. There are a range of options or dissemination streams, which vary in terms of their "safety" from confidentiality breaches. The first listed options tend to rely more on safe data whereas the last listed rely more on a safe environment, including reliance on legally binding undertakings with strong sanctions for breaches.

11. The accessibility and convenience to researchers will also vary by option.

12. Release of microdata, which is the specific subject of this session, is a key element of providing access for research purposes. Statistics legislation allows us to release microdata but only "in a manner that is not likely to enable the identification of the particular person or organisation to which it relates". Undertakings are also required. Nevertheless, there are several ways of accessing microdata whilst complying with this legal constraint. These are explained below and summarised in Table 1.

13. A Microdata Review Panel has been established to help us assess whether the disclosure risk is acceptably low (ie "not likely to enable the identification of ") for those dissemination streams that involve microdata. They look at two key risk areas:

- prevention of spontaneous identification; and
- prevention of matching risk.

14. Legal advice is that a legal undertaking preventing certain actions is consistent with "in a manner not likely to" and should be taken into consideration when making these judgements.

15. The advantages and disadvantages of each stream will be further developed in the following sections.

Table 1: Dissemination Streams to Support Research

Dissemination Stream	Notes
1. Standard Statistical Outputs	Usually in the form of tables. Restricts the type of analysis that users can undertake.
2. Datacubes	Provide more detail and the flexibility of researchers to generate their own tables.
3. Special Data Services	At the request of researchers, usually at marginal cost.
4. Confidentialised Unit Record Files (CURFs)	Data is unidentifiable. Release is on CD ROM. Equivalent to what are generally termed microdata releases.
5. Remote Access Data Laboratory (RADL)	Access to CURFs but more detailed release may be possible because of the greater control over prevention of matching with external databases.
6. ABS Site Data Laboratory	Still only provides access to unidentifiable data.
7. Collaboration	Means working through an ABS officer rather than accessing microdata directly.
8. Inhouse Analysis	In effect, working as an ABS officer working on ABS premises. This is only possible if the researcher is assisting the ABS with its functions.

16. Standard Statistical Outputs

- What does it involve? - The release of statistical outputs, usually in the form of tables, in printed and/or electronic form.

- Confidentiality Protection - This is a safe data. Standard ABS Confidentiality Practices are applied.

- Advantages - Convenient and cheap. Provides a good indication of full range of data. Increasing availability of electronic data in downloadable form improves convenience of use for further analysis. Easily accessible to a range of researchers. Low cost to the ABS.

- Disadvantages - Limits the types of analysis that can be undertaken. Not possible to undertake analysis that relies on microdata.

- Current State of Play - Increasing the availability of data in this form on the web site. Improving the availability of supporting metadata.

- When to use? - Should not be underestimated as a convenient means of supporting research. Should be a key consideration of the dissemination strategy for all statistical outputs.

17. Datacubes

- What does it involve? - The release of detailed statistical matrices that have already been confidentialised. It is a more appropriate form of release when confidentiality protection can be automated, particularly for small cells (eg population census). Special confidentiality provisions for trade data also allow data to be released in this form.

- Confidentiality Protection - This is safe data. Standard ABS Confidentiality practices apply (unless there are special provisions which exist for some data eg trade).

- Advantages - Reasonably convenient access to more detailed data than standard statistical outputs.

- Disadvantages - Same as for Standard Statistical Outputs. Also, design of good datacubes is not straightforward. Some researchers also find it difficult to use datacubes. Will not be possible to produce confidentialised datacubes for many statistical outputs.

- Current State of Play - We are slowly increasing the availability of datacubes. Increasing the knowhow of the designers of datacubes.

- When to use? - Will generally be more useful for personal data than business data. For some statistical outputs, should be considered as part of the dissemination strategy.

18. Special Data Services

- What does it involve? - The release of statistical outputs, not necessarily tables, at the request of researchers.

- Confidentiality Protection - This is safe data. It will not be provided to the researcher unless confidentiality is already protected.

- Advantages - The data and form of delivery can be tailored to the researchers need.

- Disadvantages - Will be expensive to some researchers (and for the ABS to service). Analysis limited by inability to work interactively. Researcher cannot apply own adjustments (eg for outliers) to the microdata. Turnaround to different runs of the data analysis might be slow.

- Current State of Play - Offered as a service but demand is not great. Not trying to develop, except for key clients and selected areas (eg regional statistics).

- When to use? - Usually for tabular outputs when not provided through standard outputs and access to microdata is not possible. Other forms of analysis are more likely to be run as a collaborative arrangement (see below).

19. Confidentialised Unit Record Files (CURFs)

- What does it involve? - The release of microdata files on a CD ROM which have been amended so that the identification of an individual person or organisation is unlikely.

- Confidentiality Protection - A Microdata Review Panel advises on the adjustments that are required to protect the confidentiality of the data. This may involve data amendment techniques such as deletion of some variables, reducing the detail available in some variables (particularly geography), deleting some highly identifiable individuals, and random perturbation. The confidentiality is further protected by requiring a legal undertaking from all researchers accessing the microdata. In cases of breaches sanctions will be applied (including the withdrawal of the microdata service) to the researcher and possibly their institution. Legal recourse may also be sought.

- Advantages - Great flexibility and convenience to the researcher.

- Disadvantages - Not all the detail being sought is available. Generally CURFs are not available for data about businesses. There have been a small number of breaches of the Undertaking (but not identification of individual records).

- Current State of Play - Will remain a significant dissemination stream for supporting research and secondary data analysis. Demand is high. We are trying to improve the timeliness of our releases.

- When to use? - Is regarded as one of standard outputs from household surveys. Used selectively for other surveys where data is still useful for research purposes after confidentiality protection has been applied.

20. Remote Access Data Laboratory (RADL)

- What does it involve? - Running jobs submitted by authorised users via the internet against CURFs held at the ABS, and returning analysis results after largely automated confidentiality checks. Similar to CURFs except it should be possible to provide access to more detailed data because matching risk can be controlled as data does not leave the ABS. Limited to range of analysis software provided through RADL (eg SAS, SPSS). Outputs will be manually inspected before onward release.

- Confidentiality Protection - Advice of Microdata Review Panel. Manual inspection of outputs enhanced by automatic triggers to identify output that may require rigorous inspection. Audit trails and records kept. Legal undertakings will need to be made. Sanctions against offenders.

- Advantages - Access to more detailed data. Access to analysis software that might not be available to the researcher. Free processing facilities.

- Disadvantages - Inconvenience compared with CURFs. Some delays in the release of outputs. More expensive for the ABS to administer.

- Current State of Play - Was launched in April 2003. Will be modified in light of user reaction. The number of data sets available through this facility will increase, over time.

- When to use? - Will use rather than CURF service when data matching risk of CURF is too great, and reliance on undertaking/sanctions is risky. For example, it may be used for linked data files, particularly if one of the linked files is available externally.

21. ABS Site Data Laboratory

- What does it involve? - Similar to RADL except that no downloading of unit record data is available (this is possible in RADL for up to 30 records to support outlier detection, etc). Note that it is different to situation in many other countries where a declaration of secrecy enables on-site access to unconfidentialised unit record files. We cannot do this unless the researcher is genuinely assisting the statistician to perform his functions and his employment status means that the researcher can be deemed an ABS officer. This would mean payment for services.

- Confidentiality Protection - Similar to RADL except that there is more control on output; no unit record data can leave the ABS.

- Advantages - Access to data that may not be possible through CURFs or RADL (eg longitudinal data files). More direct access to ABS experts.

- Disadvantages - Inconvenience of working on ABS premises. Expensive for ABS staff to manage, particularly across nine offices.

- Current State of Play - Is available now. Main use has been for longitudinal data files, particularly where the sample unit or some of the data has been derived from the administrative system of another agency.

- When to use? - Only when CURF or RADL service is deemed inappropriate for a data set or the researcher prefers this form of working and ABS is prepared to support.

22. Collaboration

- What does it involve? - Working collaboratively with a researcher to produce an output (often a published output) of relevance to the ABS. May or may not be a statistical output released by the ABS. The arrangements generally do not prevent researchers publishing or presenting the results of this work elsewhere, including in scientific journals.

- Confidentiality Protection - The research collaborator does not directly access unit record data. This is done by the ABS staff member working with them.

- Advantages - Mutual benefits from collaborative effort. Genuine knowledge transfer. Researcher could mostly work away from the ABS office. May result in funding being made available to the researcher to assist with research. Costs to researcher will generally be lower. Potential access to Australian Research Council grants.

- Disadvantages - No direct access to data. Limited to collaborative projects of interest to ABS.

- Current State of Play - Policy on collaborative arrangements has been put in place. Analysis Branch has been established and has been in operation for four years with about 30 staff members. This has provided a real focus for collaborative effort with the research community. Previously, arrangements were ad hoc.

- When to use? - In cases where collaboration will result in outputs of mutual benefit. For some higher priority projects, the ABS may seek collaborators. As well as confidentiality, principles that should govern collaborative work are consistency with government purchasing principles, deriving statistical value, evenhandedness and transparency, and protecting the ABS reputation.

23. In-House Analysis

- What does it involve? - The ABS can engage persons as "officers" if they are undertaking functions to support the ABS in its activities. In these situations they can access unit record data although subject to the same secrecy provisions of other ABS officers. This may be appropriate when the ABS wishes to produce an output where the researcher can cover an identified gap in expertise. Generally, arrangements can be made to allow researchers to publish aspects of their work elsewhere with permission.

- Confidentiality Protection - Secrecy provisions apply as they are ABS officers. Liable to severe penalties for breaches.

- Advantages - Provides researcher access to unit record data. Mutual benefit from collaborative work.

- Disadvantages - Much of the work will need to be done on ABS premises. Limited to subjects of direct relevance to the ABS. Some restrictions on research outputs. May not always be possible to employ as an ABS officer.

- Current State of Play - This provision has been rarely used. Recent changes to public service arrangements make it easier to implement.

- Where to use? - When the ABS takes the initiative to engage a researcher to assist it with its statistical activities. (There still may be mutual benefits of course.)

IV. LINKED DATA SETS

24. Linked data sets are a special case of a microdata set that users may want to access. Here I am talking about using data matching techniques to bring together unit records to form a set of composite records. The composite record may be based on a hard match using identifiers or a statistical match using a combination of variables (eg geography, age, sex, household characteristics). Both are of concern from the point of view of confidentiality. Hard matches are clearly of greater concern but research we have undertaken indicates a surprising high proportion of exact matches when undertaking statistical matches, particularly for files that include the household structure.

25. Linked data sets may comprise:

(a) matching ABS data sets;
(b) matching an ABS and non-ABS data set; or
(c) matching non-ABS data sets.

In (a) and (b), the ABS must be the custodian and access has to be through the dissemination streams discussed in this paper. It is not necessary for the ABS to be custodian for (c) but there are advantages. We have legislation which could underpin the arrangements for accessing these data sets and protect their confidentiality. Furthermore, our reputation is such that there is strong public confidence that we will be a trusted custodian. We also have the tools and systems to support access.

26. A linked data set can have considerable analytical power as illustrated by the following examples

- studying the interactions of a person with different institutions - by say linking together the records of health services provided by medical practitioners, hospitals, nursing homes and the like;

- studying the relationships between inputs, outputs and outcomes by drawing together information on policing, courts and prisons; and

- studying through time patterns by assembling a longitudinal database.

27. Some additional principles are needed for creating/working with these data sets. The core principles will be as follows.

- Consistency with the ABS mission to use statistical information to better support informed decision making, research and discussion.
- A demonstrable statistical benefit.
- Integrity and openness about applications.
- Publication of a statistical output from each linked data set.
- Maintaining public trust by ensuring ABS legislation, privacy legislation and other relevant legislation is followed.

28. We are considering the establishment of an Ethics Committee to help us with decisions in this area.

V. WHAT ARE OUR PLANS?

29. Until recently, the situation for each of the dissemination streams, to support external researchers, was as set out below.

- Standard Statistical Outputs - A standard service was available for all fields of statistics.
- Datacubes - Under development but available for some statistical series (eg demography, labour force).
- Special Data Services - Available but only used occasionally. Usually for production of detailed tables.
- CURFs - A regular output from household surveys, occasionally for business surveys, but needing to curtail detail released because of increasing matching risks. Also needing to strengthen the legal undertakings that are necessary for release.
- RADL - Service not available.
- ABS Site Data Laboratory - Used occasionally but not promoted.
- Collaboration - Used only occasionally in the past but over the last year or so, about forty collaborative arrangements have been established. This includes 12 collaborative arrangements as a result of the Australian Census Analytical Program.

- Inhouse Analysis - Used rarely.

30. In the future, all eight streams will be used to support external researchers. Because of their expense, we will try to limit special data services (stream 3) to key clients. Because of the increased matching risk, there will be some contraction of the detail available on CURFs. Nevertheless they will remain a key means of researcher access to microdata.

31. The key areas of development will be Standard Statistical Outputs (Stream 1), RADL (Stream 5) and Collaboration (Stream 7). We expect that more Datacubes (Stream 2) will be released but, realistically, it is only a suitable form of output for a limited range of statistical series.

32. Our objective under "Standard Statistical Outputs" will be to increase the amount of data that will be available in this form through our special web based services (eg AUSSTATS). All statistical areas will be asked to review their dissemination strategies with the view to reducing reliance on paper publications and increasing output available electronically.

33. RADL is a new service which will have just commenced operation by the time of the CES meeting. We see this as an area of further development in light of experience with the first version. It will be especially targeted at:

- providing microdata access to more detailed data sets; and

- providing access to linked data sets, especially where one of the data sets are held externally.

34. We are pursuing "collaboration" more actively now that we have a fully effective and highly respected Analysis Branch. We will attempt to initiate collaboration in these areas of greater interest to us, particularly when a new statistical output might result. Of special interest is adding value to existing data sets through analytical techniques. In practice, some researchers will approach us in the first instance. We will assess whether there are likely to be mutual benefits from collaborative arrangements. Dissemination Stream 8 may be appropriate for some collaborative projects but it is an approach we would use selectively.

VI. ORGANISATIONAL ARRANGEMENTS

35. The leadership for these arrangements must come from the ABS Executive especially whilst they are going through a period of substantial change. Communication is important, both internally and externally. We are supported by the ABS Branch responsible for policy and coordination.

36. The actual management and administration of the arrangements lies with our Information Services Division. Within this Division, they have a unit responsible for the administration and distribution of CURFs, RADL and the ABS Site Data Laboratories. They are also responsible for promoting these services and managing the relationship with clients.

37. Access methods are still under development in many respects. To strengthen our research capabilities, including research done elsewhere, and to provide more focus, we have created a Data Access and Confidentiality Methods Unit within Methodology Division. This is headed by a senior methodologist.

38. A special project team (oversighted by a Project Board) was established to support the development of RADL. The ongoing responsibility for maintenance of these systems has been transferred to Information Services Division.

39. Analysis Branch is responsible for the setting up and managing most of the collaborative arrangement that rely on access to microdata. Some may be managed through the statistical areas but this will be an exception. We are using Analysis Branch to ensure greater consistency of approach. Furthermore, they have the technical knowhow to work most effectively with research collaborators. A Project Board of our most senior subject matter statisticians oversights this work.

40. Finally, the statistical areas need to be closely involved. They are responsible for providing the underlying data sets for all the dissemination streams. Furthermore, researchers will need to call on their subject matter expertise from time to time.

VII. KEY ISSUES

41. It is becoming more and more difficult to provide truly "safe data" so it is inevitable that we will need to rely more on "safe settings", including legal arrangements, to support secondary data analysis. This is more labour intensive - requiring additional resource commitments when NSOs are often under resource pressure. Still, we believe it is an appropriate reallocation of resources if our data is being used effectively.

42. Researcher acceptance of these arrangements may be an issue. From their point of view, they may provide unnecessary constraints or inconveniences. They ask why can't we trust them to do the right thing? The communication strategy is vital. We not only need to inform the researchers of these new arrangements but why the constraints are necessary. They are much more likely to work within the system if they understand the rationale.

43. We are really moving from a paradigm of risk avoidance to risk management. There are greater risks of a loss of public confidence in the degree to which we protect the confidentiality of their data. The risks may be small, and justified by the value being added to our statistical data, but they still exist. The value system of researchers is different to that of official statisticians. The research imperative dominates and researchers can be frustrated by what they see as unnecessary impediments and bureaucracy. It is inevitable that some will "step across the line". It is unlikely that a researcher will try to identify an individual - that is not the motivation. Rather, from our experience, they are more likely to bend the rules to advance their research agenda (eg we have found cases of our microdata being on-sold to support further research albeit with added value). It is important that we act in these cases.

Legal sanctions may be appropriate in some cases. These can be difficult and drawn out. Withdrawal of service, including from the host institution, is easy to apply and very effective particularly if the message gets around the research community that the ABS is prepared to undertake this step.

44. Finally, there is a lot of international collaboration among the research community. They will point out what they can do in country A compared with country B. We know from personal experience. There would be considerable benefits if there was a greater degree of uniformity in our approaches. We have agreed on a Fundamental Principles of Official Statistics - why not fundamental principles for use of microdata by external researchers? I elaborate on this in the next section.

VIII. CONCLUSIONS

45. Supporting external research use of our statistical data is an important way of getting more value out of our statistical activities. We regard this support as an important ABS objective. Furthermore, our legislation provides us with the authority to support this type of activity.

46. In the past, we have interpreted this legislation in a conservative way - focussing on approaches that result in safe data. The increasing sophistication of technology, and the availability of external databases make it more difficult to release truly "safe" microdata (or safe datacubes for that matter). The increasing prevalence of private sector databases may be the biggest concern as there is generally less regulation about their use or misuse.

47. Consequently, there is a need to move towards dissemination approaches that rely on a "safe environment". We have been assisted in this respect by confirmation that legally binding undertakings signed by researchers can be taken into consideration when assessing whether we are complying with our enabling legislation. That is, we don't need to rely on safe data alone.

48. We will continue to take a somewhat conservative approach to interpreting our legislative authority for releasing microdata, although not as conservative as previously was the case. This is because of our concern that one significant incident could create severe damage to our reputation and our ability to maintain public confidence in the degree to which we protect the confidentiality of their data. This will affect response rates in our collections and the quality of statistics we produce.

49. The most promising new approach to providing microdata access is RADL. Its use as a means of increasing access to linked data sets is of particular interest.

50. Like many statistical endeavours there is great scope to learn off each other - both good and bad experiences. Microdata access may be an area of activity on which we may want to agree on some core principles. The research communities work across countries and make comparisons. In fact, there are already arrangements (eg Luxembourg Income Study) where microdata from several countries are brought together for convenient research access.

51. Legal and administrative arrangements will vary from country to country of course. But there still may be some core principles on which we agree. To start the debate, I suggest the following.

- It is appropriate for microdata collected for official statistical purposes to be used to support research and secondary data analysis under prescribed conditions that prevent misuse.

- The use of microdata to support external use for other than research and statistical purposes is not supported.

- There should be a legal or other arrangement to support use of microdata in order to increase public confidence in its appropriate use.

- The uses of microdata should be transparent, and publicly available, again to increase public confidence that microdata is being used appropriately.

- External researchers should not be engaged by the NSO as an employee unless they are contributing to work which will lead to an NSW output.

- The arrangements for microdata access should be cleared with the privacy authorities of the country.