AUSTRALIAN BUREAU OF STATISTICS
ANNUAL MICRODATA REPORT

INTRODUCTION

The release of microdata in the form of Confidentialised Unit Record Files (CURFs) has become an integral component of the ABS output strategy, and is a key element in achieving the ABS mission of assisting and encouraging informed decision making, research and discussion. CURFs are a very important means of providing researchers and policy analysts with the opportunity to undertake secondary analysis of ABS datasets. In 2002-03 the Statistician approved 170 CURF releases for statistical purposes.

While the release of CURFs is critical to the ABS mission, there are legal obligations imposed on the ABS in relation to the release of CURFs. As well as these, it is also important to recognise that privacy is an important community issue and the guarantee of the protection of the confidentiality of information is absolutely critical to the operations of the ABS. It is the basis on which the ABS achieves such high response rates and retains the confidence of the Australian public.

This article draws heavily on papers presented to recent meetings of the Australian Statistics Advisory Council. It details some background to the release of CURFs by the ABS, discusses issues associated with their release, and provides a summary of recent developments in providing access to microdata.

LEGISLATION

A report (No. 192) by the Joint Parliamentary Committee on Public Accounts issued in 1981 entitled 'The Collection and Dissemination of Statistics - a Discussion Paper' recommended that the statistics legislation be amended 'to allow for maximum utilisation of the data available'. One of the responses to that report was a Ministerial Determination made in February 1983 under section 13 of the **Census and Statistics Act 1905**, which amongst other powers gave the Statistician discretionary authority to release information in the form of unidentifiable individual statistical records (clause 7).

A copy of clause 7 of the Statistics Determination is provided at Attachment A. However, the key elements of the Determination are as follows:

- release of information in the form of individual statistical records requires the approval in writing of the Statistician

- all identifying information such as name and address has been removed

- the information is disclosed in a manner that is not likely to enable the identification of the particular person or organisation to which it relates
- no attempt will be made to identify particular persons or organisations
- the information will be used only for statistical purposes
- the information will not be disclosed to any other person or organisation
- the Statistician has been given a relevant undertaking.

The Determination also provides for the Statistician to impose any other condition in the undertaking that in the opinion of the Statistician is reasonably necessary. In that regard the existing undertaking prohibits matching of the file to any other file.

The main feature of the legislation is that 'the information is disclosed in a manner that is not likely to enable the identification of the particular person or organisation to which it relates', which is consistent with the ABS obligation under the **Census and Statistics Act 1905** to maintain the confidentiality of the information provided to it.

A breach of an undertaking referred to in clause 7 is an indictable offence, subject to a penalty of up to a $5,000 fine and/or two years gaol.

ABS PROCESSES

Given the sensitivities associated, the release of CURFs are tightly controlled. A Microdata Review Panel (MDRP), has been established within the ABS with the responsibility of assessing all proposals for the release of microdata prior to any submission being made to the Statistician. In undertaking the assessment the MDRP must satisfy itself that release of the CURF will meet the ABS' legislative obligation not to release information in a manner that is likely to enable the identification of a particular person or organisation. Specifically, the MDRP assesses the likelihood of identification of a particular person or organisation through spontaneous recognition (an unusual record that is known to represent an individual or organisation) or through matching with other data sources.

If the MDRP is of the view that there is a risk of identification either through spontaneous recognition or through matching, then the ABS uses a number of methods to reduce the likelihood of identification. These include:

- reducing the amount of classificatory detail, e.g. reduced geographical

information

- not releasing values as collected, but showing them as classes
- randomly perturbing the value by some small number
- swapping information between records
- dropping individual records from the file
- substituting imputed for collected information.

MATCHING RISK

The ABS is aware that the increasing sophistication in data management and information technology makes the challenge of maintaining the confidentiality of microdata more difficult. The potential to match records on CURFs with other datasets through their household demographic information, although expressly prohibited by the undertaking required by the Statistician, creates matching risks that need to be considered and addressed.

Recognising the increasing risks of disclosure arising from the existence of other datasets that might potentially be matched with a CURF, the ABS recently sought legal advice as to the extent the data presented in the CURF must be proof against a deliberate attempt to identify individuals, and the extent to which the ABS could rely on other aspects of the manner of release, such as the undertakings provided by users that they will not attempt to identify users or match the CURF against other databases.

The advice received is that the ABS is not required to consider remote possibilities or require 'absolute proof' against a determined attempt to identify individuals when considering the 'manner' of release. Additionally, the ABS is entitled to make a reasonable assumption that users will comply with the conditions of an undertaking in assessing whether disclosure is likely to enable the identification of an individual or organisation. At the same time the obligation remains on the ABS to remain alert to circumstances that might increase the risk of identification. If identification is 'likely' to happen, the ABS cannot disclose the information, or must change the manner of disclosure to remove the likelihood of identification occurring.

RECENT DEVELOPMENTS IN MICRODATA RELEASE

In light of the potential risks associated with matching and the legal advice the

ABS has received, the ABS has been reviewing its strategy for microdata release.

There are three main strands to the strategy:

i. maximise the availability of microdata to researchers and policy analysts, consistent with the legal obligations and our obligations to the community in terms of maintaining the confidentiality of information
ii. strengthening the legislation applying to the release of microdata
iii. the development of alternate modes of access to microdata.

The first of these is a reaffirmation of ABS commitment to ensuring, and where possible expanding, the continued availability of microdata to researchers and analysts. In so doing the ABS is cognisant of its legal obligations and more importantly the need to behave in a manner that will retain the confidence of the Australian public in terms of the protection of confidentiality.

In relation to the second element, the ABS is currently progressing amendments to the Determination, so as to require undertakings from each individual in an organisation accessing a CURF as well as an undertaking from the organisation as required by existing legislation. The intent is to highlight to individuals their personal obligations, and potential penalties for failure to adhere to the undertaking, as well as to ensure high level organisational commitment to the undertaking.

In respect of the third element, the ABS has now developed and has in place three distinct modes of access to ABS microdata. These are:

- CD-ROM - this is the current mode of release and where possible the ABS intends to continue to release microdata on CD-ROM in a manner similar to present release arrangements

- Remote Access Data Laboratory (RADL) - is a new development where users will be able to access ABS microdata from their desktop via secure web arrangements. The first version of the RADL was released in April

- ABS Site Data Laboratories - where users will need to come on to an ABS site to access microdata in a controlled environment.

In the not too distant future, it may be possible to create off-site data laboratories on the premises of selected clients. The data would still be held in our databases, but accessible from the remote site where it would be managed by an ABS staff member.

PROTECTION STRATEGIES

The strategy to ensure that microdata is released in a manner that is not likely to enable the identification of a particular person or organisation is applied at three points:

- the data itself is protected by providing less detail and perturbing a small number of values
- the method of access controls the amount and nature of information which can be viewed or retained, which in turn determines the amount of information potentially coming into contact with external databases
- the actions of the user are limited to those specified by the undertaking and accompanying documents.

The combination of these is different for the different methods of access, but, for all methods the combination provides the required level of protection. In all cases the undertaking prohibits the user from attempting to identify individuals or organisations from matching to unit data and from passing on the unit data to people who have not signed the ABS undertaking, and restricts use to statistical purposes. The undertaking is supported by improved educational material to facilitate users awareness and understanding of their responsibilities.

For a CURF released on CD-ROM, the level of detail and the amount of perturbation is set to achieve protection against the same identification risks as previous releases on CD-ROM, namely matching to lists of persons and spontaneous recognition. The content of CURFs on CD-ROM will be influenced by timing constraints, the content of previous releases and the external environment. The ABS is also moving towards having greater consistency in the detail provided for standard variables, both across releases and over time. As a result new releases of some CURFs on CD-ROM may have a lower level of detail compared to earlier releases. The method of access provides no additional protection and the undertaking is relied on to protect against matching using household structure.

For a CURF accessed through the RADL, the level of detail and the amount of perturbation is set to protect against spontaneous recognition. The RADL keeps the full file within the ABS computing environment, automatically limits the number of unit records which can be viewed, and limits the volume of other output that can be obtained without clearance from the ABS that it does not reveal unit data. Keeping all but a small part on the CURF within the ABS is the key protection against large scale matching to lists. All jobs and their outputs are

retained and a proportion are audited. Users are required to observe some limitations on their code, but only to the extent necessary to allow the automatic procedures to work. Users are required to keep secure any output which reveals unit data.

The level of detail for data accessed through the ABS Site Data Laboratory (ABSDL) is similar to that for data released through the RADL, but the ABSDL would be used where there was a greater risk of disclosure; for example where some user potentially had very detailed knowledge of the data.

The advantages and disadvantages for the user are different for the different methods of access. The CD-ROM allows access in the user's own environment, but to data with the least amount of detail. The RADL allows access via the users own computer, with minor limitations on code and limitations on the size of outputs which can be obtained without ABS clearance. The data in the RADL can be more detailed than on CD-ROM. Access through the ABSDL is on ABS premises, and output must be manually cleared and cannot reveal unit records.

SANCTIONS

An integral component of the protection strategies detailed above is compliance by users with the undertaking they sign. As part of the revised strategy, there will be regular audits to ensure that users, both individuals and organisations, are complying with the obligations of the undertakings. The Statistician will address any breaches of those undertakings by withdrawing the CURF service from the offending party, both individuals and organisations, and, in the case of serious offences, legal options may be pursued.

FUTURE DIRECTIONS

The ABS recognises that there is a need for all three modes of delivery (i.e. CD-ROM, RADL and ABSDL) for the foreseeable future and the ABS will continue to provide all three, however, the content of the files may vary due to the circumstances noted above.

An area that the ABS has commenced exploring is the development and availability of linked datasets. These are a special case of microdata that involve 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 (e.g. geography, age, sex and/or household characteristics). To the extent that they provide much richer datasets they are of great interest to researchers and analysts.

Linked datasets may comprise:

i. matching ABS datasets
ii. matching an ABS and non-ABS dataset
iii. matching non-ABS datasets.

In (i) and (ii), the ABS must be the custodian and access has to be through the dissemination streams discussed in this article. It is not necessary for the ABS to be custodian for (iii) but there are advantages in that the ABS has legislation which could underpin the arrangements for accessing these datasets and protect their confidentiality. Furthermore, our reputation is such that there is strong public confidence that we will be a trusted custodian. Regardless, the ABS will not undertake initiatives in this area without prior consultation with The Office of the Federal Privacy Commissioner.

Both RADL and ABSDL provide a means of increasing access to linked datasets in that while there are no limitations on what unit record or summary information a user can view within RADL or ABSDL, all information going in or out is effectively cleared by the ABS.

CONCLUSION

There is increasing demand for access to microdata to support a range of research and secondary data analysis. The objective of the ABS over the past 12 months has been to put in place the framework to facilitate and extend the availability of microdata to researchers and policy analysts. At the same time it is critical for the ABS to continue to meet its obligations to the Australian community in terms of ensuring the confidentiality of information supplied.

The development of the RADL and the strengthening of the legislation and administrative procedures have been the key to achieving that outcome. The RADL has the potential to provide a framework for significantly extending the range of microdata released in the future. At the same time there will be greater scrutiny of users to ensure that they are fulfilling their obligations, particularly as they relate to the legal undertaking they sign.

STATISTICS DETERMINATION - CLAUSE 7 Attachment A

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

(a) all identifying information such as name and address has been removed;

(b) the information is disclosed in a manner that is not likely to enable the identification of the particular person or organisation to which it relates;

and

(c) the Statistician has been given a relevant undertaking for the purposes of this clause.

(2) A reference in paragraph (1) (c) to a relevant undertaking shall be read as a reference to an undertaking in writing by:

(a) in the case of information to be disclosed to a person, being an individual - that person;

(b) in the case of information to be disclosed to an official body - the responsible Minister in relation to, or a responsible officer of, that official body; or

(c) in the case of information to be disclosed to an organisation other than an official body - a responsible officer of that organisation;

that use of the information will be subject to the following terms and conditions:

(d) no attempt will be made to identify particular persons or organisations; and

(e) the information will be used only for statistical purposes; and

(f) the information will not be disclosed to any other person or organisation; and

(g) if the Statistician considers it necessary in a particular case - either or both of the following:

(i) that the information, and all copies (if any) of the information, will be returned to the Statistician as soon as the statistical purposes for which it was disclosed have been achieved;

(ii) that access by officers to information, documents or premises will be given as may be necessary for the purpose of conducting a compliance audit concerning observance of the terms and conditions under which the information is disclosed; and

(h) any other condition that, in the opinion of the Statistician, is reasonably necessary in a particular case.'