Research Access to Microdata: an attempt to provide a context

*by Angela Dale **

Background

National statistical institutes have an obligation to compile statistics that provide the information required by government. In the UK, following the 1980 review by Sir Derek Rayner, the remit of the Government Statistical Service was restricted to meet the specific needs of government departments rather than the broader needs of the business community, local government and academia. However, the launch of National Statistics in June 2000 involved an explicit commitment to meet the needs of a broader range of users that included the general public. The Framework Document (June 2000) that accompanied the launch set out the Government's commitment to providing a "statistical service that is open and responsive to society's needs and the public agenda: better and more reliable official statistics that command public confidence." Under the *Aims and Objectives of National Statistics*[1], in section 3, the third bullet point lists:

> To provide researchers, analysts and other customers with a statistical service that assists their work and studies;

However, statistical offices have to tread a careful balance between providing the data needed by all sections of society and maintaining the confidence of the general public who supply most of the data. The experience of some other countries shows that if the public lose confidence in the national statistical office then the process of data collection will be undermined and may not recover. For example, Germany has not taken a full population census since the census planned for 1983 had to be postponed until 1987 because of public concern over proposals to use census returns to update the local population registers. The Netherlands has not taken a census since 1971, following a significant level of refusal in the 1971 Census and poor test results in 1979.

Data in the public domain

In the past the general public only had access to government statistics through reports in local libraries. However, in recent years greater dissemination by statistical offices, largely through the opportunities offered by the web, have brought statistical information into the homes of a large sector of the population and into the offices of voluntary organisations, schools and other locally based organisations. Particularly through the development of Neighbourhood Statistics (which includes data from the 2001 census, surveys and also administrative sources), there is now readily accessible information about the places where people live and work. In addition, there is also unrestricted on-line access to reports on the social and economic conditions of the population and the tabulations that underpin them – for example, the Living in Britain Report produced annually by the Office for National Statistics (ONS).

There is, therefore, a developing reciprocal relationship between the population that provides the data and the statistical office which collects and compiles that data. For the first time the average person in the street, or student in school (as well as businesses and local authorities) is able to obtain recent and high quality data from the UK statistical offices without charge. The very high rate of hits on the ONS web-site, and Neighbourhood Statistics in particular, suggests that the public are, indeed, accessing these data. This development should be an important step towards retaining and increasing public acceptance of the conduct of the census and government surveys.

However, the very fact that these data are public and easily available means that they must not reveal any identifiable information, either now or at some unforeseen time in the future. But it is next to impossible to predict what technologies or techniques may become available in the future that could lead to the identification of individuals and what motivations there may be for using them. Therefore the balance between providing a public service by making data easily available and ensuring the confidentiality of the data is very difficult to get right.

An additional and little-researched factor is the impact of public perceptions. A wrong belief that people can be identified in government statistics may be as damaging to public confidence as the reality – and, for many people, the two may not be distinguished. What evidence is available [1] suggests that people are unsure about the extent to which information they supply in a census is passed to

other government departments - and this is also the case in the USA and Australia. There is also confusion about the source of information used in direct marketing and whether or not it comes from government data sources.

Protecting confidentiality
It is widely accepted that geographical detail is a key factor in identifying individuals. In small geographical areas (e.g. the 2001 census output areas with about 125 households) residents are likely to have good knowledge of the characteristics of their neighbours. In this size of area there may only be one woman aged 45 who is living in privately rented accommodation or only one man of Black Caribbean origin who works in education. To ensure that such an individual cannot be identified, much less detail on characteristics such as industry, occupation, age and ethnic group can be provided for small geographical areas than for larger areas. This is reflected in census outputs, when tables at local district level have more detail than those at the level of output area. In addition, ONS have added protection to tables from the 2001 census that have small cell sizes. Cells containing 0, 1, 2 or 3 respondents have been changed to 0 or 3.

For many members of the public and many researchers, information about local areas is what is required. Where information on national or regional social and demographic characteristics is needed then tables are available on a range of topics.

The role of academic research in the social sciences
However, for many researchers these publicly available data sources provide only a first port of call. Academic research needs to go beyond published reports and pre-prepared tables to conduct original research using microdata (that is, individual records for individuals and households).

Academic social research has a vital role to play in understanding social change. It can provide methodologically rigorous analysis of issues that are of fundamental importance: for example the household composition of the ageing population, migration patterns, ethnic diversity, regional differentiation and much more. Academic analysis can go beyond the descriptive to seek explanation and to test hypotheses. Multivariate analysis is needed that includes all variables of importance to the outcome of interest. Furthermore, these variables need to be derived in a way appropriate to the analysis. For example appropriate age groupings will vary by whether one is analysing labour market activity or family formation. Bespoke classifications need to be developed that are specific to a particular analysis – for example measures of exclusion based on information about all household members. Existing classifications or indicators need to be subject to challenge and to re-working based on different definitions. At the heart of scientific research is the requirement that results are published and open to challenge. The ability to replicate analyses is fundamental to good scientific practice.

It is also essential that data collected at public expense is used as extensively as possible, consistent with the undertakings given to the respondents. In this spirit, the results of research should be available in an accessible and reader-friendly form as well as through publications in scientific journals.

Analysis of microdata files from the 1991 UK Census has had a major research impact, including analyses of unemployment that allow both individual and area-level characteristics to be included [2] and analysis of ethnic differences in women's employment over the life course [3]. A summary of this research is available from the CCSR web site (www.ccsr.ac.uk/sars/findings).

However, there is an increased risk of identification with microdata by comparison with pre-defined tables, and this is recognised in the procedures used to ensure that confidentiality is protected. The first protection is that microdata files represent only a sample of the population. Therefore there is only a small chance – perhaps 2 or 3 in 100 - that an individual will be included. In addition, care is taken over the amount of detail that can be released and geographical detail is always heavily restricted. Finding the appropriate balance requires careful assessment of the risk of data disclosure. But it also requires recognition that **absolute** safety jeopardises any significant research activity. Therefore the risk of **not** supporting research also has to be considered. It is also worth noting that, where breaches of confidentiality have occurred, (see above) these have not been associated with research use of data.

Safety: a double balancing act
We can define two interacting dimensions when considering access to data - the level of safety associated with the dataset; and the level of safety associated with the access setting.

Level of safety associated with dataset
This will depend heavily on the degree of detail in the data; the proportion of the population in the sample; the ease of identifying the data either through matching or spontaneous recognition. Thus a microdata file with a low level of risk may be a sample with very restricted individual detail and little geographical information. Level of risk will also vary with the extent to which disclosure protection methods (e.g. perturbation or data swapping) have been used on the data.

Level of safety associated with access setting
This will range from access confined to a safe setting within the statistical office – at one extreme – to unrestricted access where data is distributed to users with few if any conditions of use.

The two dimensions interact so that, at one extreme, if the data are judged to be entirely safe, then the access arrangements can be very open. This is exemplified by the Public Use Microdata Files produced by the US Bureau of the Census, which can be downloaded without restriction from the web-site of the US Bureau of the Census. These files are samples – 1% and 5% - where the amount of both individual detail and geographical information has been heavily restricted to preserve confidentiality.

By contrast, if the data are very detailed and/or contain information that could be used to identify someone, then greater safety needs to be built into the access conditions. An example is the ONS Longitudinal Study that contains data with a great deal of individual and geographical detail, from the census and from vital events, but where access is highly restricted and only available within a secure setting inside ONS.

We have, therefore, a continuum from safe data to safe setting – with all protection built into the data in the former and all protection built into the setting in the latter.

Research and safety

Public use microdata files are of considerable value because they can be readily used anywhere at any time. Access is quick and easy and these kinds of data are ideal for teaching, where students need to interact with data. However, datasets that are safe enough to need no restrictions will usually lack some of the detail required by researchers. For example, in safe data variables such as occupation or ethnic group may be very broadly banded and thus may not provide the distinction required for some analysis purposes. A lack of geographical detail may also hamper research into the respective effects of individual characteristics and local labour markets. Some bias may also have been introduced into the data through perturbation or suppression, in order to ensure that unusual individuals or households cannot be recognised. These are all concerns which have been addressed in the development of microdata samples from the 2001 Census.

At the other end of the spectrum, secure in-house access, e.g. within ONS, where researcher credentials are screened, all data is available under strictly controlled conditions and all outputs are carefully checked, can allow access to much more detailed data. In this kind of safe setting the analyst may be able to access detailed geographical information on place of residence or place of work, or data that is very sensitive – for example information on cause of death, cancer registration or, in the case of business surveys, information on business performance. However, in-house safe settings are expensive to set up and run and also difficult for researchers who have to travel long distances and spend considerable time away from home.

Finding the middle ground

The two extremes of safe data and safe setting both have disadvantages for conducting research. We therefore need to explore a range of options that lie between these polar opposites and that can allow researchers access to data that is of sufficient detail and quality to meet research needs while also retaining the level of confidentiality required by the national statistical institutes.

Fundamental to this middle ground is the need to recognise that researchers have no interest in breaching confidentiality. Research is concerned with establishing statistically significant differences between social and demographic groups, not with attempting to identify individuals. Researchers do, however, have a very strong interest in promoting good practice and respect for research data.

The safeguards set out below provide varying degrees of protection and can be used singly or together to increase data protection beyond that required for public use files. They should therefore allow a concomitant increase in detail in the data.

The role of institutional controls

Research is conducted in recognised institutions (one definition of a research institution is recognition to administer research grants). These institutions can be asked to accept responsibility for research data used by their staff. This control was used in the UK with dissemination of the Samples of Anonymised Records from the 1991 Census. Institutions where staff or students wanted to use the data were asked to identify a responsible person who actively managed data access.

Microdata under licence

Statistics Netherlands provides access to microdata for research purposes under license. Researchers in the UK who wish to use microdata from the Data Archive are required to agree to a confidentiality undertaking. However, this could be extended to provide a more explicit and binding contract between the researcher and the statistical office. This would include use of the data for a fixed length of time and a requirement to return all copies of the data after that time.

A safe setting on-site

In Canada and the USA, statistical offices are increasingly setting up secure data centres for analysis of microdata files. These represent safe settings that, for the researchers who happen to be located nearby, can provide access to the most detailed microdata. Whilst these settings can provide very safe conditions, they are expensive to run and privilege those able to use the facility. Nonetheless, it is possible to imagine a situation where most universities could support a safe room that would allow access to relatively detailed microdata. There are established

procedures for access controls to prevent data being removed from the room. This should be far enough along the safe setting spectrum to allow access to much more detailed microdata than that released as public use files.

Increased use of technological developments
There are a growing number of examples of remote safe-settings where microdata files are held on a secure server that may be located in a statistical office or any other safe location. Access to the data can be indirect – as with the Luxembourg Income Study, where the researcher submits a request to run an analysis; the request is physically downloaded and moved across a firewall to a secure server holding the data. The results, which are controlled to prevent disclosure, are then returned by email and the researcher has no access to the actual microdata. Alternatively, researchers may be able to interrogate data files through the use of additional controls such as a password authorisation system backed up by a license agreement and registered IP addresses for authorised computers. Increasingly, the Grid and associated middleware allow imaginative solutions that can maximise research use whilst retaining confidentiality.

Conclusions
In a time of increased concern over data security there is a growing need to explore all possible ways in which data collected at public expense can be fully analysed, while at the same time ensuring the confidentiality of the respondents. In the spirit of ensuring that there is some payback to the public who provide responses to censuses and surveys, there is a strong argument that accessible research findings should be posted on national statistics web-sites. By doing so, we would make the value of research based on government data more apparent to all.

[1]Fieldhouse, E. and Gould, M. I. (1998) "Ethnic minority unemployment and local labour market conditions in Great Britain," Environment & Planning A 30, No.5, 833-53.

[2]Framework For National Statistics, June 2000, http://www.statistics.gov.uk/about_ns/downloads/FrameDoc1.pdf

[3]Holdsworth, C. and Dale, A. (1997) "Ethnic Differences in Women's Employment," Work, Employment and Society 11, 435-57.

[4]Marsh, C. (1993) "Privacy, confidentiality and anonymity in the 1991 Census" in Dale, A. and Marsh, C. (eds) The 1991 Census User's Guide, London: HMS

Notes

[1] Section 3 of the ONS Framework Document

http://www.statistics.gov.uk/about_ns/downloads/FrameDoc1.pdf

An earlier version of this paper was published in Significance, issue 1, of the Royal Statistical Society We are grateful for permission to reproduce it here

I am grateful to Chuck Humphrey, University of Alberta and Joris Nobel, Statistics Netherlands for helpful comments and suggestions.

* Angela Dale, CCSR, University of Manchester. Contact: Angela.Dale@man.ac.uk