GSS Disclosure Control policy for Microdata Produced from Social Surveys

Contents

1. Purpose

The UK Statistics Authority has issued its Code of Practice for Official Statistics (CoP), in which Principle 5: gives the requirements for confidentiality[1]. This GSS disclosure control policy for the release of microdata derived from Social Surveys provides guidance to ensure compliance with Principle 5 of the CoP, in particular where social survey microdata are lodged at the UK Data Archive for access under the End User Licence. The governance of the release of microdata by ONS is described as an example of good practise.

The Statistics and Registration Service Act 2007 (SRSA) includes data confidentiality regulations which apply to ONS. This policy therefore also contains guidance for ONS to ensure compliance with the SRSA.

The following topics are included in this paper and more details are provided in the Appendices:

- Definition of different categories of microdata
- Implications for release of microdata of the Code of Practice for Official Statistics
- Statistics and Registration Service Act
- Procedures for making microdata available to users
- Understanding the key characteristics of the data and outputs
- Assessing disclosure risk
- Reconciling user requirements with the need for disclosure control
- Legal and policy considerations
- Disclosure control methods

2. Scope

Social survey microdata are files consisting of individual records for each respondent. The records include demographic details about the respondent together with variables specific to the survey. These data provide a valuable research tool for a wide range of users and uses. Therefore, in accordance with Principle 1 of the UK Statistics Authority Code of Practice (CoP), microdata files of different degrees of disclosiveness are made available to other government departments and to academic and other researchers.

This GSS Policy provides guidance on releasing microdata in accordance with the CoP and specifically Principle 5 (see appendix 3). The policy only applies to microdata derived from social surveys. Guidance for other types of microdata from other sources will be developed in the future. In addition the policy provides guidance for ONS to ensure compliance with Section 39 of the SRSA, see 3.2.

Social survey microdata may be lodged at the UK Data Archive (UKDA) for secondary research access under an End-User Licence (EUL) see section 3.3. They may also be provided to Eurostat for the use of researchers and analysts across the European Union under equivalent arrangements to the EUL. In addition, the ONS may release microdata to customers under licence, see 4.1 (2).

This policy, and particularly Appendix 1, focuses on the provision of data under the EUL. However section 4 provides guidance on classifying data into different degrees of confidentiality, and how these may be handled.

3. Context

3.1 The UK Statistics Authority Code of Practice for Official Statistics

The Code of Practice (CoP), which is published on the web-site www.statisticsauthority.gov.uk, sets out the professional principles and standards which official statisticians are expected to follow and uphold. In particular Principle 5 deals with the data confidentiality of private information[1].

[1] Principle 5: Confidentiality. Private information about individual persons (including bodies corporate) compiled in the production of official statistics is confidential, and should be used for statistical purposes only. (See Appendix 3 for the full text of this principle.)

Data which are not private information are defined as those which do not reveal the identity of an individual or organisation, nor any private information relating to them, taking into account other relevant sources of information. We refer to such identification as disclosure, and thus data are private information if there is a risk of disclosure. This is discussed further in 3.4.

This GSS Policy is in line with the CoP, and includes guidance in the preparation of microdata which are not private information, and may therefore be released for research purposes in accordance with Principle 5, for example by being lodged at the UKDA under the EUL, see 3.3.

3.2 The Statistics and Registration Service Act (SRSA)

The SRSA, which established the UK Statistics Authority, came into force on 1 April 2008. Section 39 of the Act deals with the confidentiality of personal information held by the Authority, and applies to data held by the ONS as its agent and to any recipient of ONS data. Thus ONS has to take account of both the CoP and the SRSA when releasing data. Section 39 specifies what constitutes a disclosure of information and the sanctions that may apply for any breach of confidentiality. The full text of section 39 can be found in Appendix 4.

Data which are personal information are defined as those which could reveal the identity of an individual or organisation, or any private information relating to them, through being specified in the information, by being deduced from the information, or by being deduced from the information when taken together with any other published information (section 39 (3)). In order to be able to provide research access to EUL level data (see 3.3), the data must not be personal information.

The SRSA recognises, however, that some valuable research will require access to personal information. Therefore there is an exemption in Section 39 allowing applications to be made under the Approved Researcher gateway for access to more detailed datasets. (Such data were previously available under the UKDA Special Licence; use of this term henceforward should be understood as meaning data which are only available to Approved Researchers.) ONS has published the processes relevant to Approved Researchers on its web-site[2], see "Access to ONS data service". The ONS procedures for releasing personal information are described in section 5.1, below.

3.3 The UK Data Archive End-User Licence

The UK Data Archive (UKDA), www.data-archive.ac.uk, is a repository of data for the use of national and international social researchers. It is provided by the Economic and Social Data Service (ESDS). Government departments regularly lodge data at the UKDA in order to make it available to researchers. Registered users may access datasets under an End User Licence (EUL).

The UKDA has the facility to restrict access to certain data to authorised users only. ONS makes use of this facility by depositing "special licence" datasets, access to which require ONS authorisation and is restricted to Approved Researchers.

The terms and conditions of the EUL provide a certain level of confidentiality protection[3]. Practice 1 of Principle 5 says that data must be protected by taking into account other relevant sources of information, see 3.1. The conditions of the EUL mean that the only other relevant sources of information which need to be considered are those in the public domain. A user who attempted to identify an individual from EUL data by means of a private source of information would be in breach of the Licence and subject to penalty. The full text of the EUL is attached as Appendix 5.

3.4 Preventing disclosure of private/personal information

Social survey microdata are based on statistical units, which may be individual survey respondents, households, families or other bodies, such as schools or employers. This policy provides guidance on

---

[2] http://www.ons.gov.uk/about/who-we-are/our-services/unpublished-data/access-to-ons-data-service/index.html

[3] The UKDA end-user licence includes the following restrictions:

- The user must preserve the confidentiality of individuals and households in the data.
- The user must not attempt to derive information from the data relating to an individual, nor claim to have done so.
- The data must be kept secure and may only be shared with other registered users of the UKDA.
- The data may only be used for research or educational purposes, and may not be used for commercial purposes without permission.

avoiding disclosive situations, thus protecting the statistical units which make up the data. An individual or organisation which might seek to discover the identity of a statistical unit is referred to as an intruder. To ensure that private information is sufficiently protected, we consider scenarios which might make disclosure possible.

For EUL microdata, as a minimum, the following scenarios need to be considered:

- Using published datasets, together with the microdata, to identify an individual or a household.
- Spontaneous recognition, where an intruder recognises an individual or a household in the microdata from published information.

The scenarios indicate which variables in the microdata might make them private/personal information, and these can then be protected. Ways of protecting microdata include recoding (i.e. banding of some variables), suppression, perturbation, randomisation and imputation. Of these, recoding and suppression impose the least burden on data providers, and are therefore generally recommended. More detail is provided in Appendix 1.

3.5 Criteria for preparing microdata for release under the EUL

When preparing EUL microdata from a social survey there are three distinct criteria:

- The resulting microdata must be protected such that an intruder would not be able to identify an individual, family or household, either directly from the data or by using information in the public domain.
- The microdata need to include enough detail to meet the requirements of the majority of users.
- The disclosure protection process must not impose an unreasonable burden on the business areas.

Any solution needs to achieve compliance with these three conditions and guidance is given in Appendix 1.

4. Procedures for different categories of microdata

4.1 Categories of microdata

The SRSA, CoP and EUL allow microdata derived from social surveys to be categorised into three tiers of information according to their level of "identifiability":

(1) Private/personal information

Private information as defined by the CoP, and personal information as defined by the SRSA, are data from which it is possible to identify an individual, or private information relating to him or her, either directly from the data or by the auxiliary use of published information. Examples of personal information derived from ONS social surveys are:

- identified data, which are made available via the Virtual Microdata Laboratory (VML),
- GSS datasets, which are anonymised data supplied to other Government departments,
- ONS "Special Licence" datasets, which have had some degree of protection applied, and which are made available to Approved Researchers (see 3.3).

(2) End-User Licence datasets.

These are not private information under the CoP nor personal information under the SRSA, because there is no risk of disclosure resulting from the use of data in the public domain. They are lodged at the UKDA under the End-User Licence (EUL), which is considered to provide adequate protection against disclosure resulting from the use of other data, such as private databases. The End-User Licence is attached as Appendix 5. If required, these data may also be supplied to Eurostat.

The ONS also provides some datasets of this type to customers under a licence similar to the EUL, for example LFS data.

Appendix 1 provides guidance on how to design these datasets.

(3) Public use data

These data are neither private information under the CoP, nor personal information under the SRSA. Account must be taken of the risk of disclosure from data in the public domain and private sources of information since the data are released without any kind of licence arrangement. However, in order to protect microdata to this level, the utility would be seriously reduced. The ONS does not currently release any microdata as public datasets.

4.2 Restrictions on release of different categories of microdata

(1) Private/personal information

The SRSA, permits ONS to release personal information under one of the exemptions in Section 39 (4), see above, 3.2. Examples are provision of data which are required under separate legislation, or release of data to an "Approved Researcher".

Under the CoP, private information may be released where this has been authorised by the National Statistician or the chief Statistician in a Devolved Administration (Principle 5, practice 5). The CoP states that confidential information should only be used by trained staff who have signed an appropriate declaration. Also when confidential data are supplied to a third party this must be authorised and recorded, and there must be written confidentiality agreements in place. (See Appendix 3.) This GSS Policy offers the ONS Microdata Release Panel procedure as an example of best practice for compliance with these requirements of Principle 5 (see section 5).

(2) End-User Licence datasets.

Data may be released under the End-User Licence provided sufficient disclosure protection has been applied to ensure that they do not reveal private/personal information, taking account of the possible use of data in the public domain. ONS policy is that data providers should ensure that it would take a disproportionate amount of time, effort and expertise[4] for an intruder, using published information, to identify a statistical unit to others, or to reveal information about that unit not already in the public domain. In the case of Social Survey data, the statistical units are generally individual respondents and households. This guideline is recommended to the GSS.

The GSS microdata policy makes a clear distinction between EUL microdata and public use data. Principle 5 of the CoP requires that official statistics should not disclose private information, taking into account other relevant sources of information. Thus if data were to be released for public use, data providers would need to take into account the possibility that an intruder might have access to a private data source or to privileged information which could be matched with the microdata to enable the identification of an individual.

Since EUL datasets need only take precautions against the use of data in the public domain, there is a possibility that an individual could be identified from EUL microdata by an intruder using a private data source or privileged information. As discussed in 3.3, it is considered that the End-User Licence provides adequate management of this risk since the conditions of the EUL, see Appendix 5, provide some control over the use of data.

(3) Public use data

The only public use data currently released by ONS are tables and other statistical representations which are not personal information. Tables and other statistics derived from social survey data are subject to the GSS Disclosure Control Policy for Tables Produced from Surveys.

[4] The designer should allow for the intruder to have access to powerful data processing software and hardware equivalent in standard to that available in ONS, to have some statistical and mathematical expertise equivalent in standard to those found in an ONS Statistical Officer and to be prepared to dedicate a number of hours of their time to the task of identifying an individual.

5. Implementation

This section describes the procedure for microdata release which is used by ONS. It is recommended as an example of best practice, and ONS can advise other members of the GSS who wish to put similar processes in place (see 5.2).

5.1 ONS procedure for releasing microdata

The ONS procedure for releasing microdata is facilitated and controlled by the Microdata Release Panel (MRP) and the Microdata Release Unit (MRU). These are part of ONS Legal Services. The MRP is made up of senior representatives of different business areas within ONS. A data provider wishing to release microdata outside of ONS submits an application to the MRP describing the data, the purpose for which it is being released and any related data security and confidentiality procedures, including a data access agreement (DAA). The MRU has developed a range of templates for DAAs to cover different circumstances.

Applications to release datasets which are not personal information, such as EUL microdata, must include a risk assessment from the Statistical Disclosure Control (SDC) team, in ONS Methodology. The Microdata Release Unit is authorised by the MRP to approve releases of non-personal microdata, subject to a satisfactory assessment by SDC. However applications to release datasets which are personal information must be submitted to the full Panel for approval.

5.2 Service provided by the SDC centre

When an ONS business area or data provider wishes to release data which are not personal information, a risk assessment must be obtained from the SDC team. To help in making this assessment, the data provider is asked to complete a checklist, see Appendix 6. This allows the presence of key variables to be specified, together with information about any way in which they are protected, e.g. by banding. The risk assessment, which may include advice on further protection needed, is attached to the MRP application.

Within ONS, the SDC team can provide advice to the data provider on how to ensure that the data are not personal information. However it is the responsibility of the data provider to ensure that, when microdata are made available to the UKDA under the EUL, SDC advice has been fully implemented. It is also the data provider's responsibility to ensure that the data released correspond to the data specification provided in the MRP application, which is the definitive record of the data provided to the customer. The specification on the MRP application may be used for subsequent releases of the dataset and must therefore be accurate.

GSS members wishing to obtain assistance from ONS should contact the SDC team, SDC@ons.gov.uk. Depending on the amount of work involved, it may be necessary to discuss charging for it under the ONS Methodology Consultancy Service arrangements.

5.3 Documentation

Where disclosure control has been applied to microdata, corresponding documentation should be supplied to the UKDA or other user, so that researchers will be aware of the methods used and any subsequent impact on their analyses.

5.4 Reviewing and developing the process in ONS

Methods of risk assessment are continually being reviewed and developed by the SDC team. There is ongoing work to monitor what data are publicly available, and this will inform the intruder scenarios which need to be considered when carrying out disclosure risk assessments. This guidance may therefore be subject to future amendments.

6. Responsibilities

The Responsible Statistician is responsible for confidentiality protection of released data, ensuring that the standard disclosure control methods are applied, and that any other special circumstances are taken into account. This role will always and only be taken by the National Statistician, the Chief Statistician in a devolved administration or a head of profession. Day to day management of disclosure control for data

release may be delegated to output managers, data managers or others responsible for the confidentiality guarantee pertaining to the social survey, whether the data are released by GSS or by others using data from this source.

Within ONS, the National Statistician has delegated responsibility for the release of microdata from social surveys to the Microdata Release Panel. The MRP takes advice from various relevant areas within ONS, including the SDC team, the business areas, and security and legal units.

Appendix 1 - General guidance for producing social survey microdata which are neither private nor personal information

A1. Introduction and definitions

This Appendix provides general guidance for ensuring that microdata derived from social surveys are neither private information as defined by the CoP nor personal information as defined by the SRSA. It starts by defining some terms used, then discusses ways to assess the disclosure risk of data and makes recommendations on disclosure control methods. Examples are based on ONS social surveys. ONS uses this guidance for microdata to be released to the UKDA under the End-User Licence (EUL).

The procedure which ONS uses to regulate release of microdata under the EUL, or an equivalent licence, is described in section 5.1. This is recommended as an example of good practice.

A1.1 Definitions

Identification Data which identify an individual, not necessarily by knowing his name, which may be used in order to learn or record something about that individual, or where the processing of the data has an impact upon that individual.
Intruder Someone who deliberately or inadvertently determines confidential information about a respondent, or attempts to do so.
Scenario A situation which allows an intruder to gain such information.
Direct identifier A variable which directly identities an individual. Examples are name, address, National Insurance number, NHS number.
Key variables Variables which, when taken on their own or in combination, may assist an intruder and allow a respondent to be identified. These are generally indirect identifiers. Examples are age, occupation, household composition.
Identification Key A combination of key variables
Sample unique A record for which the identification key is unique within a dataset.
Population unique A record for which the identification key is unique within the underlying population.
Disclosure risk The probability that an identification made by an intruder is correct.

A2. Risk assessment

The risk assessment of microdata needs to take account of several factors, including possible intruder scenarios, key variables, size of sample and the presence of large households.

A2.1 Scenarios

When assessing the disclosure risk of microdata, it is necessary to first consider the intruder scenarios. These are assumptions about what an intruder might know about respondents and what information will be available to match against the microdata and potentially make an identification. The consideration of scenarios indicates some of the variables which are likely to be used by an intruder.

The definitions of scenarios and the corresponding key variables were developed by Professor Angela Dale and Dr Mark Elliot. (Elliot and Dale, 1998; Elliot and Dale, 1999). Two of the scenarios which they defined were the commercial database cross-match and spontaneous recognition. For EUL microdata, these are the main scenarios which need to be considered and they are described further below. Other scenarios may also need to be considered, either because of particular characteristics of the survey, or as we become aware of more publicly available data.

A 2.1.1 **Scenario 1 – use of published datasets**

An intruder in possession of published datasets can use key variables to match these against the microdata. This may enable him to identify an individual.

Examples of such published information are:
- The Electoral Register.
- Commercial datasets, such as consumer profile databases, which may be purchased at a reasonable cost by any member of the public.

The Confidentiality and Privacy Research Issues Group (CAPRI), at the University of Manchester, undertook a Scoping Study for a Data Environment Analysis Service (DEAS) (Purdam and Elliot 2006). This found that a large number of databases are held by commercial data companies, which combine public records, such as the electoral register, with other sources, such as lifestyle surveys, to produce large datasets which are then made available commercially to the public. As the number of such datasets increases, potential intruders will be able to make use of multiple datasets and thus enhance the level of information derived from them.

Example of a published dataset

A typical example of a commercially available dataset is one produced by CACI Ltd. This is a hierarchical dataset. For every household included in the dataset there is a record for each adult in the household, and each of these records contains variables which give details, such as age group and sex, of every child in the household. Thus records can be grouped into households. See A2.4 for further discussion of hierarchical microdata.

Records in this dataset include the following variables:

- Name
- Address
- Postcode
- Age
- Sex
- Ethnicity
- Number of cars
- Number of children, given in 5 year age-groups (e.g. 1 child aged 0-4, 2 children aged 5-9)
- Size of household
- Tenure
- House type
- Number of rooms
- Occupation (high-level)
- Income (banded)
- Qualifications

Thus Scenario 1 assumes that an intruder would link published information with EUL microdata using an identification key comprising demographic variables. They would then have the direct identifiers, name and address, linked with all the other information in the EUL dataset. There would be a high probability of these being correct matches.

A2.1.2 **Scenario 2 – spontaneous recognition**

An intruder may spontaneously recognise an individual in the microdata by means of published information. This can occur for instance when a respondent has unusual characteristics and is either an acquaintance or a well-known public figure such as a politician, an entertainer or a very successful business person. An example is the “Rich List” which publishes annual salaries of high-earning individuals.

The key variables for this scenario include:

- Name
- Age
- Sex
- Marital status
- Income
- Occupation – job title, which may be equivalent to 4-digit occupation and industry codings

A2.2 Key variables

Taking into account these scenarios, and others if appropriate, we can identify which variables an intruder is likely to combine into an identification key. Such a key could be used to attempt to match the microdata with other published data to which the intruder has access, in order to identify one or more records as referring to particular individuals. The disclosure risk comes from individuals within the microdata that are both sample uniques and population uniques on the key since this increases the probability of a match and

therefore re-identification. Provided that the disclosure risk is reasonably small, the data may be considered not to be personal information; this takes account of the "disproportionate effort" rule (see 4.2 (2)).

It follows that the variables which are most likely to need protection include demographic indicators, such as geography, household composition, ethnicity, occupation etc. Salaries and household income are also key variables. Work in the USA (Winkler, 1999) showed that the level of matching between datasets is improved when amount of income data is used as part of a matching key, due to the availability of administrative tax records. Although tax data are not publicly available in the UK, some income-related data are in the public domain. Examples are the salaries of company directors, and very high salaries and bonuses, which are all published.

A2.3 Sample size

Microdata based on a larger sample will have a greater absolute disclosure risk than a smaller sample, since the number of re-identifications is likely to be greater. The larger the sample size in a set of microdata, the greater confidence an intruder can establish in a possible identification, in that the larger sample size reduces the likelihood of a similar individual existing outside the sample. Therefore microdata based on larger samples should be treated as more risky.

Suggestions of how to apply disclosure control to microdata from samples of different sizes are given below, see A3.2.1 and Table 1.

A2.4 Household surveys

Many social surveys are household-based, examples being the Family Resource Survey (FRS) and the Labour Force Survey (LFS). Microdata from these surveys are hierarchical, as they include a record for each individual in the household as well as variables which allow the individuals' records to be linked. This enables an intruder to enhance identification keys, for example by combining age, sex, marital status and the relationship of each individual in the household. Such keys increase the likelihood of households being identified. Thus the disclosure risk has to be assessed at the individual and household level. Suggestions for addressing this factor are given below in A3.2.4.

A2.5 Large households

The presence of large households increases disclosure risk. It has been demonstrated that households of size eight and above are intrinsically disclosive, independent of the size of the sample (Elliot, 2005). Work on the Sample of Anonymised Records from the 2001 Census (2001 SARs) noted that, for private households of size 6 and above in England, 88% were population uniques for age-sex structure. (Bycroft et al, 2005.) Suggestions for dealing with large households are given below in A4.

A2.6 Longitudinal data

Some surveys use the same respondents for a number of successive periods. For instance four waves of respondents may be used, with each wave contributing to four successive surveys and being replaced in turn over the four periods. Examples of ONS surveys which use such methodology are the Labour Force Survey (LFS), where each wave takes part in four successive quarterly surveys, and the General Household Survey (GHS), where each wave takes part in four successive annual surveys.

ONS is also developing new longitudinal household surveys, such as the Wealth and Assets Survey which will use the same panel each year.

Microdata from such surveys have increased disclosure risk, as successive datasets may be combined to assist in identifying a contributing household or individual. For some longitudinal surveys it is likely that there will be a requirement for EUL microdata to permit users to link sample members across successive years. Suggestions for longitudinal datasets are given below in A3.2.6.

A3. Disclosure control

Having considered the likely intruder scenarios and identified the risk factors, the process of preparing microdata which is not personal information can be divided into three steps:

(1) Anonymise the data
(2) Apply disclosure control to key variables
(3) Deal with large households

A3.1 Anonymising the data

This means removing all direct identifiers, including name, address, post-code, NI number and NHS number. If the data contain any other direct identifiers, such as Passport Number, then these must also be removed. In addition date of birth must be removed, and it is generally advised that all similar variables such as year of birth and month of birth should also be removed.

Anonymised data can still be personal information.

A3.2 Applying disclosure control to key variables

As discussed above, the data provider should consider the relevant intruder scenarios, what key variables are included in the data and the size of the sample. The following advice is based on experience gained from disclosure risk assessments which the Statistical Disclosure Control (SDC) team has carried out on the various sets of ONS social survey microdata, and in particular work carried out on the Sample of Anonymised Records from Census 2001 (Gross et al, 2005). The advice may be changed for future microdata releases if the SDC team becomes aware of relevant published information which could be used by an intruder (see 5.4 and A 2.1.1).

Table 1 is a list of some key variables with suggestions of ways they may be protected. These are suggestions and not rules, and the list is not exhaustive. The method of disclosure control chosen should be appropriate for the survey and the sample size. Users' requirements should always be borne in mind; if a variable is needed at a lower level than advised, then another variable should be protected at a higher level. For example, if exact rather than banded age is required, then salary and income variables could be banded instead, or occupation and industry variables provided at a higher level. Where appropriate, reference is made to the scenarios described above as Sc1 and Sc2.

A3.2.1 Size of sample

The following is a rough guide on sample sizes:

Small - If the sample is less than 1% of the population, then most key variables may not need to be protected. However there will always be some, such as geography, which need treatment, see suggestions below.

Medium - If the sample size is between 1% and 3% of the population, then it is likely that several key variables will need to be protected.

Large - If the sample size is greater than 3% of the population, then further protection may be necessary. However no social surveys, whether current or envisaged, belong to this category, so the guidance in this appendix only refers to small and medium sized surveys.

A3.2.2 A note on geography

Geography variables are the primary candidates for protection, as removing low levels of geography introduces extra uncertainty into a possible identification. For most EUL microdata, the lowest level of geography is therefore Government Office Region (GOR). One of the main reasons for setting up the Special Licence, see 3.1, was to give researchers access to data with lower geographical details, such as local authority. Some variables, such as the urban/rural indicator, are based on postcode, so their inclusion may reveal a lower level of geography. Care needs to be taken of variables such as Council Tax and associated variables. Because local authorities publish their rates of council tax, this could reveal the local authority.

However there will be some surveys for which it is appropriate to include local authority or unitary authority, for example where the primary purpose of the survey is to look at local matters. In such cases, care will need to be taken that the level of detail of other variables is correspondingly reduced. For example ages could be banded into 10-year age groups, salaries could be banded, and information on variables such as family structure, number of children, etc, could be reduced. It is also important that variables which are based on postcode, such as urban/rural indicator and deprivation factor, are not included in such EUL

datasets. Note that, if variables such as deprivation factor are also essential for researchers, they can sometimes be represented by quintile or decile values. Each case needs to be considered on its own merit, but if LA/UA is included it is essential that no lower geography can be deduced from the data.

As stated in section 5.2, the SDC team in ONS Methodology can be consulted if help or advice is needed.

A3.2.3 Examples of protecting key variables

Table 1 is not exhaustive and is provided only as a guide; data providers are the experts on their data and will therefore be aware of variables which may pose a risk. Care should be taken that, when a variable is protected, all variables derived from it are similarly protected.

Table 1 – examples of protecting key variables to make data non-disclosive

Key variable	Reason for protection	Suggested treatment
Geography – respondent's residence, place of work etc	See A3.2.2 and Sc1	Lowest geographical level should generally be GOR, (or Wales as a country).
Age – respondent's age, age left full-time education, age of oldest child in household under 16, etc.	Key variable in Sc1 and Sc2	Small samples: single year of age may be provided. Medium size samples: ages should be banded, into say 5 year groups. (see A3.2.1)
Size of household	Key variable in Sc1	See A4, below.
Country of birth/nationality	There are around 144 possible values (and increasing) of these variables.	Consider whether this level is needed. It may be acceptable to band these variables, e.g. UK, EU, other.
Occupation/industry – main job, secondary job, previous job, etc.	Key variable in Sc1 and Sc2. Coding frames for these are generally to 3 digits or 4 digits. The 4-digit level can be very disclosive in some circumstances[5].	Consider whether 4-digit level needs to be included. Recommendation is that if industry is given to 4 digits, occupation should only be given to 3 digits, etc.
Salary – gross & net, annual, weekly, hourly etc. Bonuses etc.	Company directors' salaries are in public domain. Very high salaries and bonuses are often published.	Very high salaries and bonuses should be protected by top-coding at an appropriate level. Weekly and hourly rates will need to be correspondingly top-coded. See A3.2.4
Income – household income, gross & net, etc	Key variable in Sc1.	These should be rounded to nearest £1. Very high values should be top-coded at an appropriate level, similarly to salaries. See A3.2.4
Other financial variables	These should all be considered. Examples are large winnings on Football Pools, National Lottery etc. which may be published. Council Tax rates are	Large winnings should be top-coded at £500,000 (based on 2008 values). For variables based on Council Tax, see A3.2.4

[5] If both occupation and industry are given to 4 digits, then the combination may be disclosive, e.g. company director of a particular manufacturer in that region. When SIC codes are revised, then if data have been published with low-level SIC codes, they should not be re-published with the new codes.

Table 1 – examples of protecting key variables to make data non-disclosive		
Key variable	Reason for protection	Suggested treatment
	published by local authorities, see 3.2.2.	
Number of rooms, bedrooms, etc	Key variable in Sc1	These should be based on top-coding number of bedrooms at 6, and number of rooms at 10 (excluding kitchen, bathroom).
Number of cars	Key variable in Sc1	Number of cars and vans available to the household should be top-coded at 3.
Urban/rural, Index of multiple deprivation, ACORN code	These variables should not be present in EUL datasets, as they can all disclose lower geographies.	Simple 2-value urban/rural indicator will usually be acceptable. 8-value U/R indicator should not be included in EUL datasets.

A3.2.4 Financial variables

As shown in Table 1, financial variables such as income, salary and Council Tax need extra protection.

Incomes and salaries

High incomes and salaries may be top-coded, for example at 10*average-salary in the sample, with related variables being treated similarly. An alternative method has been developed by the ONS business area responsible for the Expenditure and Food Survey (EFS):

(1) Variables are grouped, e.g. all those relating to salary, all those relating to income tax payments, etc.
(2) Within each group a key variable is identified, e.g. income tax by PAYE.
(3) The cut-off for the top 4% values of this key variable is found.
(4) This cut-off value is then used to top-code all the other variables in the same group.

The EFS method has the advantage of affecting a relatively small number of records. It is implemented as part of the programming which processes the Blaise data.

Council Tax

Some social surveys include Council Tax band for the respondent and also amount of Council Tax paid. There are also several variables based on the Council Tax payments.

Each local authority publishes their rates of Council Tax. Therefore the band plus the amount paid can lead to disclosure of a respondent's local authority (LA). It is strongly recommended that EUL microdata should not include any geographical data below Government Office Region (GOR), see 3.2.2. Therefore if both band and amount paid are included in the data, then there needs to be disclosure control of these.

There are different ways in which Council Tax payments and variables derived from them may be protected. The following method was developed by the EFS team:

(1) In each GOR, LAs are grouped according to the level of Council Tax for Band A properties
(2) For each group the average Council Tax rates are calculated.
(3) These averages replace the original value in the data.

This method means that the Council Tax variables are close to the original values, but uncertainty as to the Local Authority is introduced.

A3.2.5 Other key variables

As stated above, Table 1 is not exhaustive. Other variables such as tenure, ethnicity, number of children in household, marital status, qualifications, etc. are included in one or both of the scenarios discussed in A2.1. These need to be considered in the context of the survey and the likely requirements of users. They may be candidates for protection when dealing with large households, see A4.

Examples of variables which need care are marital status, sexual identity, case number and free text.

Marital status

Civil partnerships have introduced possible new values to the marital status variables. Civil Partnerships are discoverable data, and are therefore considered to be in the public domain.

As an example, inspection of a particular social survey dataset found that there were 2 respondents who stated that they were separated from their partner in a Civil Partnership, and 2 who had been in a Civil Partnership which had been legally dissolved . These small numbers made the microdata potentially disclosive. SDC agreed that such values should be grouped together, to give a new value for the marital status variable of "civil partner or former civil partner". This solution is recommended for all social surveys with variables including or derived from marital status.

It is likely that, as the number of people in Civil Partnerships increases, the number of people in former Civil Partnerships will also increase. This could mean that the risk of disclosure falls to an acceptable level. Thus each set of microdata should be considered on its own merits. A useful guide is that there should be at least 3 respondents in a GOR for whom the variable has the same value.

Sexual identity

This question is being introduced into some ONS surveys. The possible values of this variable are: heterosexual, homosexual, bisexual, other, and "prefer not to say". The disclosure risk here is that the numbers of bisexual and others will be very small.

In theory this would not be a problem unless data on sexual identity has been published, and could be taken together with EUL microdata to assist identification of an individual. This is unlikely to happen. But it is also necessary to consider the risk of self-identification. For example a researcher, knowing that their household was a respondent to the survey, might recognise themselves from the data, and thus discover the sexual identity of another member of their own household. SDC does not normally advise protection against self-identification, as it would make it very difficult to release EUL datasets. In this case, however, the sensitivity of the variable makes it advisable to address this risk. SDC therefore recommends the following:

- For EUL microdata, the sexual identity variable should not be included.
- For “Special Licence” datasets, i.e. datasets provided to Approved Researchers, the variable should be recoded into 3 groups; heterosexual, homosexual, and other.

Case number

The case numbers in social survey microdata can reveal information about the geography of a household. Data providers should consider whether case numbers need to be included in EUL datasets. It is recommended that case numbers should be pseudonymised.

Free text

Free text variables should never be included in EUL microdata. These include, for example, specific qualifications and job titles or other descriptive information related to employment.

A3.2.6 Longitudinal surveys

Longitudinal surveys, or surveys with a longitudinal element, such as the LFS and GHS, pose a further risk, as linking successive waves or years can disclose more information about respondents than a snapshot survey, see A2.6. Microdata from such surveys may need to be subjected to further restrictions so that successive datasets cannot be linked, as this would increase disclosure risk to an unacceptable level. Possible methods are to have a higher level of banding on demographic variables such as marital status and number of children.

If there is a requirement to allow EUL microdata to reflect the longitudinal nature of a survey, by allowing individual households to be linked over time, then additional modification of the data is necessary. For example a change in household size or marital status may allow an intruder to identify a household or individual by means of births, deaths, marriage, civil partnership and divorce registrations, which are in the public domain. Data providers should consider methods such as banding ages, marital status, socio-economic and relationship variables.

If users require more details, so that they can study the change over time in respondents' circumstances for instance, then consideration should be given to supplying the data under the Approved Researcher protocol.

A4. Dealing with large households.

As discussed in A2.4, where the survey is household based, the microdata is hierarchical and identification keys can be composed of variables such as the age and sex structure of the household and relationships of individuals to each other. In the light of scenario 1 (published datasets) this increases disclosure risk and the probability of identification.

Large households generally contain both adults and children, and at the present time there are no published datasets which include detailed information for children (see A2.1.2). Where large households consist of only adults, the risk of matching with published data is likely to be mitigated by their increased mobility. Therefore no additional protection is recommended for households of size less than 10.

However, there are very few households of size 10 and above. Based on the 2001 SARS they account for less than 0.05% of households and fewer than 0.2% of individuals. Data about these are very disclosive and the recommendation is that all records pertaining to such households should be suppressed.

It is possible that future development of commercially available datasets will increase the likelihood of being able to match them with records for large households of size less than ten. The SDC team will therefore keep this situation under review.

A5. Summary

This appendix provides guidance on preparing microdata which are not disclosive taking into account data in the public domain. They may therefore be released under an EUL as being neither private information under the CoP nor personal data under the SRSA. ONS procedure requires an application to release such data to include a risk assessment from the SDC team confirming that the data are not personal information.

Consideration of possible intruder scenarios helps to indicate what level of disclosure control is required in order to ensure that data are not personal information. Table 1 suggests some ways of protecting particular variables, but this is not exhaustive and does not cover all special circumstances. In addition, attention needs to be paid to large households, longitudinal datasets and sensitive variables. Data providers should use their knowledge both of the data and of the requirements of users to arrive at a data specification which is not personal information but retains as much utility as possible. The SDC team are able to give advice in individual cases.

Appendix 2 – References

Elliot, M. J., and Dale, A. (1998) Disclosure risk for microdata: Workpackage DM1.1 What is a key variable? *Report to the European Union ESP/204 62/DG III*

Elliot, M. J., and Dale, A. (1999) Scenarios of attack: The data intruder's perspective on statistical disclosure risk. *Netherlands Official Statistics. Vol 14, Spring 1999, 6-10.*

Purdam, K., Elliot, M. (2006) Data Environment Analysis Service Scoping Study Final Report.

Elliot, M. (2006) Assessment of disclosure risk for hierarchical microdata files.

Bycroft, C., Clift-Matthews, M., Spicer, K., Jackson, P. J. (2005) 2001 Household Sample of Anonymised Records (SAR), a report to the ONS Data Stewardship Working Group.

Winkler, W. (1999) Re-identification methods for evaluating the confidentiality of analytically valid microdata, *Research in Official Statistics, 1(2), 87-104.*

Gross, B., Guiblin, P., Merrett, K. (2004) Risk Assessment of the Individual Sample of Anonymised Records (SAR) from the 2001 Census.

UK Statistics Authority Code of Practice http://www.statisticsauthority.gov.uk/assessment/code-of-practice/index.html

Access to ONS data service http://www.ons.gov.uk/about/who-we-are/our-services/unpublished-data/access-to-ons-data-service/index.html

ONS Methodology Consultancy Service http://www.statistics.gov.uk/about/data/methodology/consultancy-service.asp

UK Data Archive End User Licence http://www.data-archive.ac.uk/aandp/access/licence.asp

Code of Practice for Official Statistics, Principle 5: Confidentiality

Private information about individual persons (including bodies corporate) compiled in the production of official statistics is confidential, and should be used for statistical purposes only.

Practices

1. Ensure that official statistics do not reveal the identity of an individual or organisation, or any private information relating to them, taking into account other relevant sources of information.

2. Keep confidential information secure. Only permit its use by trained staff who have signed a declaration covering their obligations under this Code.

3. Inform respondents to statistical surveys and censuses how confidentiality will be protected.

4. Ensure that arrangements for confidentiality protection are sufficient to protect the privacy of individual information, but not so restrictive as to limit unduly the practical utility of official statistics. Publish details of such arrangements.

5. Seek prior authorisation from the National Statistician or Chief Statistician in a Devolved Administration for any exceptions, required by law or thought to be in the public interest, to the principle of confidentiality protection. Publish details of such authorisations.

6. In every case where confidential statistical records are exchanged for statistical purposes with a third party, prepare written confidentiality protection agreements covering the requirements under this Code. Keep an operational record to detail the manner and purpose of the processing.

Appendix 4 - Section 39 of the Statistics and Registration Service Act (2007)

39 Confidentiality of personal information

(1) Subject to this section, personal information held by the Board in relation to the exercise of any of its functions must not be disclosed by—

(a) any member or employee of the Board,

(b) a member of any committee of the Board, or

(c) any other person who has received it directly or indirectly from the Board.

(2) In this Part "personal information" means information which relates to and identifies a particular person (including a body corporate); but it does not include information about the internal administrative arrangements of the Board (whether relating to its members, employees or other persons).

(3) For the purposes of subsection (2) information identifies a particular person if the identity of that person—

(a) is specified in the information,

(b) can be deduced from the information, or

(c) can be deduced from the information taken together with any other published information.

(4) Subsection (1) does not apply to a disclosure which—

(a) is required or permitted by any enactment,

(b) is required by a Community obligation,

(c) is necessary for the purpose of enabling or assisting the Board to exercise any of its functions,

(d) has already lawfully been made available to the public,

(e) is made in pursuance of an order of a court,

(f) is made for the purposes of a criminal investigation or criminal proceedings (whether or not in the United Kingdom),

(g) is made, in the interests of national security, to an Intelligence Service,

(h) is made with the consent of the person to whom it relates, or

(i) is made to an approved researcher.

(5) For the purposes of subsection (4)(i), "approved researcher" means an individual to whom the Board has granted access, for the purposes of statistical research, to personal information held by it.

(6) The Board is from time to time to publish criteria by reference to which it will determine whether to grant access as specified in subsection (5).

(7) Those criteria must require the Board to consider—

(a) whether the individual is a fit and proper person, and

(b) the purpose for which access is requested.

(8) The Board may not grant access to an individual as specified in subsection (5) unless he has first signed a declaration, in such form as the Board may determine, that he understands the requirements of this section.

(9) A person who contravenes subsection (1) is guilty of an offence and liable—

(a) on conviction on indictment, to imprisonment for a term not exceeding two years, or to a fine, or both;

(b) on summary conviction, to imprisonment for a term not exceeding twelve months, or to a fine not exceeding the statutory maximum, or both.

(10) Subsection (9) does not apply where the individual making the disclosure reasonably believes—

(a) in the case of information which is personal information by virtue of subsection (3)(a), that the identity of the person to whom it relates is not specified in the information,

(b) in the case of information which is personal information by virtue of subsection (3)(b), that the identity of that person cannot be deduced from the information, or

(c) in the case of information which is personal information by virtue of subsection (3)(c), that the identity of that person cannot be deduced from the information taken together with any other published information.

(11) In the application of this section —

(a) in England and Wales, in relation to an offence committed before the commencement of section 154(1) of the Criminal Justice Act 2003 (c. 44),

(b) in Scotland, until the commencement of section 45(1) of the Criminal Proceedings etc. (Reform) (Scotland) Act 2007 (asp 6), or

(c) in Northern Ireland,

the reference in subsection (9)(b) to twelve months is to be read as a reference to three months.

Appendix 5 ONS Methodology Group – Statistical Disclosure Control Centre

Microdata questionnaire

(Note that this document is protected before it is sent to data providers to be completed. Most reply boxes have drop down menus. In this version the default response or the first item in the drop-down menu appears in blue.)

Introduction

Where microdata are to be released as non-disclosive, they should be assessed as such by the Statistical Disclosure Control centre (SDC). The SDC assessment is then attached to the MRP application at question 10. Guidance for preparing non-disclosive data can be found in the Standards and Guidance database: ONS Disclosure Control Policy for the Release of Microdata Derived from Social Surveys.

In order to assist the SDC in ascertaining whether data are non-disclosive, the business area is asked to complete this questionnaire. Please provide as much information as possible – the more you provide the less likely it is that the SDC will need to return for further information and therefore the quicker will be the MRP process.

The following should be borne in mind when completing the questionnaire:

- Help for each text box is provided in the status bar (at the bottom of the screen). This should be referred to for each item.
- You are advised to use tab (rather than a mouse) to navigate through the form.
- Anonymised data is not the same as non-disclosive data.
- The checklist refers to the specific dataset requested, or to be released – not to the original source data.
- The business area has a thorough knowledge of its own data, and should therefore be able to highlight potentially disclosive variables, even if these are not covered by the questionnaire. Where you give details of variables in the text boxes provided, please remember that the names of variables may not indicate what they are, so please give description as well as name.
- The data definition (record layout) of the microdata should be finally determined before the SDC is asked for an assessment.
- Please do not hesitate to contact the SDC for any further assistance you may need.

When the questionnaire is complete, please send it to the SDC – currently to Carole Abrahams.

General information

Please give your name.	
What is the name of the survey or sample? If not one of those in the drop-down box, please give the full name, not just initials.	**APS**
What type of dataset?	**Individual**
What time period(s) are covered by the microdata?	**Month**
What geographical area is covered by the microdata?	

What fraction of the population is included in the sample? (Please state as a percentage)	
What sampling design was used?	
What survey design was used?	
What were the sampling units?	
Have the data been treated for outliers, and if so what methods were used?	**Outliers detected and treated (please add details)**
Have any of these data been released previously?	**Yes, to government departments**
Have similar data (i.e. these microdata but with different geographical level or different codings) been released previously?	**Yes (please give details)**
Have any tables derived from these data been published?	**Yes**
Have any bespoke tables been released?	**Yes**
Does the dataset contain a variable, such as the record id, which could be used to link it to other published datasets? This	**Yes, please give details**

includes data lodged at the UKDA under an end-user licence, datasets released for previous time periods, etc. Please also consider the possibility of linking a household dataset to a personal dataset.	

How many households of size 10 and over were included in the survey?	
Have all records for individuals in these households been removed from the dataset	Yes
Which key variables have been protected? Please give full details.	

Geographic variables

Please give details of any of the following variables included in the microdata requested (see bottom of screen for help):

Respondent's residence			
Country (please give details)		Health authority (please give details)	
Sub-national areas (please give details)		Commercial (if present, please give details)	
Post-codes (if present, please give details)		Is Council Tax band present?	**Yes**
Is amount of Council Tax present in data?	**Yes**	If Council Tax is present, has it been protected? If so please give details.	
Any other (please give details)			

Respondent's place of work	
Country (please give details)	
Sub-national areas (please give details)	
Any other (please give details)	

Are you aware of any possible geographical overlaps between the data requested and other published data? If so, please give details in the space below.

Variables relating to individuals

Please consider all variables pertaining to individuals, indicating whether or not they are present. Where they are present, please give details, including any coding. See bottom of screen for help with each item.

Age (if present, please give details)	**Single year of age**	Date of birth variables	**Exact date of birth**
Sex	**Present**	Marital status (if present, please give details, including coding frame)	**Present (please give details of cod**
Ethnicity (if present, please give details, including the number of categories)	**Present (please give details)**	Country of birth (if present, please give details including the number of possible values)	**UK, EU, non-EU**
Socio-economic	**NS-SEC, 3 class version**	Occupation (if present please state coding frame)	**SOC 2000, 1 digit**
Industry (if present please state coding frame)	**Present, please give details**	Occupation/ industry for previous/secondary employment	**Same coding as main employment**
Number of hours worked	**Present, please give details**	Salary (gross or net)	**Exact salary**
Qualifications (please give details)	**Only highest qualification only**	Age left full-time education	**Present, please give details**

In the space below please give details of any other personal variables included in the data requested, such as health information, country where qualification was obtained, number of children, etc.

Variables relating to households

Please consider all variables which might identify a particular household or family. Where they are present in the data requested, please give details, including any coding. See bottom of screen for help with each item.

Size of household (please give details)	**No. of people in h'hold, please give details**	Family type	**Present, please give detai**
Number of rooms (please give details)	**Number of rooms, please give details**	Number of cars/vans (please give details)	**Present**
Type of accommodation (please give details)	**Present**	Tenure (please give details)	**Present**
Whether respondent is HRP	**Present**	Relationships	**Relationship to HRP**
Age variables	**Present, please give details**	Household income (gross or net)	**Exact income**
Is there more than one income variable (please give details)?	**Yes, please give details**	Can families or h'holds be (re)constructed from the data?	**Yes, please give details**

Other

Are there any other variables which could possibly enable the household, or one or more respondents, to be identified?	**Yes, please give details**

Sensitive variables

EUL data should not contain very sensitive variables, such as sexual identity, HIV status, criminal record etc. Such variables should only be included in "Special Licence" datasets. If in doubt, please contact SDC for advice. Please list below any sensitive variable which have been removed from the data.