Government Statistical Service

GSS/GSR Disclosure Control Guidance for Microdata Produced from Social Surveys

October 2014

The last few years has seen continued interest from users and producers in ensuring official statistics reach their full utility whilst ensuring the risk of disclosure is minimised. A number of guidance documents are available to the GSS under the following headings.

Microdata produced from social surveys
Tables produced from administrative data
Tables produced from survey data.
These documents are all published on both the GSS and ONS websites.

As a result of this continued interest the Statistical Suppliers and User Group, a forum for engagement between producers and users of official statistics, commissioned the GSS to update its existing disclosure control guidance published in 2007. This revised guidance on disclosure control for administrative sources has been updated to reflect the views expressed by a range of users and producers. The previous version of this guidance was approved by the GSS and GSR and hence applies to both social researchers and government statisticians.

The release of microdata (record level data) is more common these days due to both the desire of researchers to access the most detailed available data and the Government lead Open Data policy which encourages departments to make data publically available.

As with tabular data there are confidentiality issues when releasing microdata. Producers of such statistics must ensure that their statistics meet the needs of users by enabling relevant analysis to be carried out while at the same time protecting confidentiality. This guidance describes the approach that data providers should follow when producing microdata.

Relevant associated documents are

The *Code of Practice for Official Statistics* See https://gss.civilservice.gov.uk/wp-content/uploads/2012/12/Official-Statistics-Code-of-Practice.pdf (CoP) and specifically Principle 5: Confidentiality,

National Statistician's guidance: Confidentiality of Official Statistics See https://gss.civilservice.gov.uk/wp-content/uploads/2012/12/Confidentiality-of-Official-Statistics-National-Statisticians-Guidance.pdf (CoOS guidance) sets out principles for how to protect personal data from being disclosed.

The *ICO Anonymisation Code of Practice*
http://www.ico.gov.uk/for_organisations/data_protection/topic_guides/anonymisation.aspx was published in 2012 by the Information Commissioner's Office and this provides considerable support and guidance to data providers charged with generating and publishing tables (or microdata) from source microdata.

The CoP states 'Ensure that arrangements for confidentiality are sufficient to protect the privacy of individual information, but not so restrictive as to limit unduly the practical utility of official statistics.' This emphasises the two requirements of statistical disclosure control. Detail relating to individual statistical units is to be protected but the released data must still be of high practical utility for users.

This document will be updated regularly to accommodate the latest thinking in statistical disclosure control. It is hoped it will be followed by a document on guidance on microdata produced from administrative data sources.

GSS /GSR Disclosure Control guidance for Microdata Produced from Social Surveys

Contents

1. Purpose

GSS / GSR statistical disclosure control guidance documents for the release of tabular outputs (both administrative and sample survey data) have previously been produced. This related document is specific to the release of microdata (record level data with each row usually defining characteristics of an individual or household) and will be of particular interest to researchers who require access to record level data along with survey managers who produce microdata. It is published on the GSS and ONS websites allowing all potential users to access the guidance.

The UK Statistics Authority has issued an overview of the Code of Practice for Official Statistics (CoP)[1], (including guidance for reporting breaches). Principle 5 of the CoP gives the requirements for confidentiality[2]. This GSS disclosure control guidance 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 (EUL). There are other methods of release and access, which are briefly discussed, such as data released under a Special Licence to an Approved Researcher but the main focus is on EUL data.

The governance of the release of microdata by ONS is described as an example of good practice. This involves all record level outputs being assessed by a Microdata Release Panel (MRP). For most departments a panel may not be a feasible approach but assessment should be carried out by at least one expert, not involved in the release process, with some knowledge of the data.

The Statistics and Registration Service Act 2007 (SRSA)[3] includes data confidentiality regulations which apply to ONS. This guidance therefore also contains guidance for ONS to ensure compliance with the SRSA. Data released by government departments other than the ONS are also considered with reference to the Data Protection Act.

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

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

- Definitions of key terms used in this document can be found in Appendix A

The Code of Practice states that data providers should ‘ensure that arrangements for confidentiality are sufficient to protect the privacy of individual information, but not so restrictive as to limit unduly the practical utility of official statistics.’ This emphasises the two requirements of statistical disclosure control. Detail relating to individual statistical units is to be protected but the released microdata must be of high practical utility for users.

[1] http://www.statisticsauthority.gov.uk/assessment/code-of-practice/index.html

[2] A summary states that ’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 C for the full text of this principle.)

[3] http://www.legislation.gov.uk/ukpga/2007/18/contents

The original guidance was approved by the GSS and GSR and hence applies to both social researchers and government statisticians. This update is distributed for comment prior to formal approval.

2. Scope

Social survey microdata consist of individual records for each household and/or individual 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 CoP, microdata files of varying degrees of disclosiveness are made available to other government departments and to academic and other researchers.

These GSS guidelines provide guidance on releasing microdata in accordance with the CoP and specifically Principle 5 (see appendix C). The guidance only applies to microdata derived from social surveys and not administrative data. In addition the following guidance will ensure that the ONS is compliant with Section 39 of the SRSA. See section 3.2 for background to SRSA in general and section 39 in particular.

Specific legislation such as the Data Protection Act (DPA)[4] is especially relevant when data are released by departments other than the ONS, although it is also applicable to ONS data. This Act specifies a number of data protection principles for users of the data such as the data should be used fairly and lawfully and for limited and specifically stated purposes. Certain personal variables, such as ethnic background, are defined as being 'sensitive' and these have stronger legal protection.

Social survey microdata may be lodged at the UK Data Service formerly the UK Data Archive (UKDA) for secondary research access under a EUL or a Special Licence (SL) for Approved Researchers[5]. There is more detail in section 3.3 on the UKDA and Data Service. Microdata may also be provided to Eurostat for the use of researchers and analysts across the European Union under equivalent arrangements to the EUL Section 4 contains background on the different categories of microdata and how they may be accessed.

This guidance and particularly Appendix A 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.

2.1 ONS Guidance documents

This short section places this document in the correct relationship to other ONS SDC guidance documents and gives some advice on which sections will of particular use to readers with different needs for both ONS and non ONS releases.

- This document applies to microdata derived from social surveys.
- Microdata derived from business surveys are usually accessed through the ONS Virtual Microdata Laboratory (VML) where all outputs are disclosure checked prior to release. This is distinct from the case here where licensing arrangements are used.
- There is additional ONS guidance for tabular outputs from both social and business surveys and from administrative data
- Specialist guidance for cancer microdata is available[6] as part of the general advice for health statistics
- Guidance for births and deaths data has recently been updated.[7]

[4] http://www.legislation.gov.uk/ukpga/1998/29/contents
[5] http://ukdataservice.ac.uk/deposit-data/how-to.aspx
[6] http://www.ons.gov.uk/ons/guide-method/best-practice/disclosure-control-of-health-statistics/index.html

Survey managers dealing with an initial request for microdata should concentrate on Section 3 onwards. More experienced managers may already be familiar with Section 3.
If the data are ONS data sections 5.1 to 5.4 are particularly useful. For non ONS releases look at section 5.5.
If the request for microdata is similar to a previous request, Sections 4, 5 and 6 are most applicable.
For a routine review of current procedures, Sections 5 and 6 should be the starting points.

3. Background

Current Standards and Laws relating to confidentiality are discussed here. Please be aware that departmental advice on relevant legislation should always be obtained prior to releasing microdata. Some of the information below refers only to ONS outputs.

3.1 The UK Statistics Authority Code of Practice for Official Statistics

The CoP sets out the professional principles and standards which official statisticians are expected to follow and uphold. In particular Principle 5 of the code considers the data confidentiality of private information.

Private information is defined in the National Statistician's Guidance; Confidentiality of Official Statistics[8] as being information that:

- relates to an identifiable legal[9] or natural person, and
- is not in the public domain or common knowledge, and
- if disclosed would cause them damage, harm or distress

When assessing if microdata reveal private information it is necessary to take into account other relevant sources of information that might in combination reveal private information. These could include additional datasets held by a researcher such as membership lists of particular pressure groups.

The Guidance states that producers of official statistics should be aware of the expectation individuals may have when their information is used to produce statistics. Information relating to an individual should be considered by a producer of statistics to be 'private' if it was provided with the expectation that the information would be kept out of the public domain. Survey pledges provide respondents with assurances that the information they provide will remain confidential.

These GSS guidelines are in line with the CoP, and include 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 UK Data Service for access by researchers under the EUL. See section 3.3 for more detail.

3.2 The Statistics and Registration Service Act 2007 (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

[7] http://www.ons.gov.uk/ons/guide-method/best-practice/disclosure-control-policy-for-birth-and-death-statistics/index.html
[8] http://www.statisticsauthority.gov.uk/national-statistician/ns-reports--reviews-and-guidance/national-statistician-s-guidance/index.html
[9] A 'legal person' is a company, enterprise, or other organisation that has a legal identity. A 'natural person' is a member of the public. Where the term 'individual' is used in the Code it means both legal and natural persons, both living and dead.

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 of the SRSA 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 D. The SRSA is applicable to the ONS and other Government departments will have different legal requirements to meet.

In the context of SRSA Section 39 personal information is defined as information which relates to and identifies a particular person (including a body corporate). 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 (SRSA section 39 (3)). In order to be able to provide research access to EUL level data (see section 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 produced by the ONS. The ONS procedures for releasing personal information in microdata are described in section 5.1.

As noted above the SRSA is applicable only to ONS data. Other government departments will have to consider legislation relating to datasets they produce in combination with their own specialist knowledge of the data when deciding on the disclosure control required before publishing microdata. One such Act is the previously mentioned Data Protection Act which defines certain variables as being 'sensitive data'. This Act is applicable to all published data, ONS outputs have to follow the DPA as well as the SRSA. There is more detail on the DPA in A3.2.5.

3.3 The UK Data Service

The UK Data Service http://ukdataservice.ac.uk/ is a repository of key economic and social data for the use of national and international researchers. It is funded by the Economic and Social Research Council (ESRC) with contributions from the Universities of Manchester and Essex and replaced the UK Data Archive in 2012. As noted at the start of Section 2 they have recently published new guidance on depositing data. Government departments regularly lodge datasets at the UK Data Service in order to make available to researchers. Registered users may access datasets under an EUL or a SL where more detailed data are available under more restrictive access conditions.

The terms and conditions of the EUL provide a certain level of confidentiality protection[10]. Practice 1 of Principle 5 says that data must be protected against disclosure having taken into account other relevant sources of information, see section 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 F.

Access to SL data requires approval of an additional application justifying why access to variables released only under SL are required by the user. Since the advent of the SRSA access to ONS SL datasets is restricted to Approved Researchers. A prospective user must apply to ONS to become an Approved Researcher in order to access an ONS SL dataset. Non-ONS datasets released by the UK Data Service, such as the Welsh Health Survey, require the explicit permission of the data owners for

[10] The UK Data Service end-user licence includes the following restrictions:
- The user must preserve the confidentiality of individuals and households and organisations in the data.
- The user must ensure that the means of access to the data (such as passwords) are kept secure and not disclosed to anyone else
- The data must not be used for commercial purposes without obtaining permission

release to specified researchers. Other government departments will need to determine if this is a suitable method of releasing SL datasets.

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 document provides guidance on avoiding disclosive situations, thus protecting the statistical units which make up the data. An intruder is the term given to an individual who identifies or attempts to identify a respondent in the microdata. Attempts at identification can range from a malicious attempt to undermine the credibility of the organisation releasing the data to spontaneous recognition of an individual from viewing the data. To ensure that private information is sufficiently protected, we consider various scenarios which might make disclosure possible.
For microdata to be released under EUL, 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.
- Identification of an individual or household in longitudinal data through changes in attributes over time that form an identifying pattern.

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, post 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 A.

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 other 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 producers of statistics

EUL microdata cannot be released as a Public Use File because this would suggest that the data are non-personal and would be subject to the Freedom of Information Act (FOI) (see Section 4.2).
Any solution needs to meet these three conditions and comply with the guidance as given in Appendix A. There is always an element of risk when releasing EUL microdata. However the aim is that any assessment carried out prior to release will reduce the risk level to a low level while maintaining a relatively high level of utility.

4. Guidance for different categories of microdata

Microdata from social surveys can be categorised according to whether they disclose private/personal information, or would do so in combination with other data sources.

4.1 Microdata that disclose 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 them, either directly from the data or by the auxiliary use of published information. The Data Protection Act as mentioned in Section 3.2 defines particular variables as sensitive personal data, more details in A3.2.5. Examples of the release of personal information derived from ONS social surveys are:

- identified data, which are made available via the VML to Approved Researchers (see section 3.3) subject to these researchers having received the relevant training modules. Detailed documentation on using the VML is being updated but there is some background on approved researcher accreditation[11].

- GSS datasets, which are anonymised data supplied to other Government departments, An example is Labour Force Survey data which was supplied to the Department for Education (DfE) in 2010 without names but with a range of identifying variables.

- ONS SL datasets, which have had some degree of protection applied, and which are made available to Approved Researchers.

The SRSA (see section 3.2 above) permits ONS to release personal information under one of the nine exemptions in Section 39 (4),
These exemptions occur when the information

- is required or permitted by any enactment,
- is required by a Community obligation,
- is necessary for the purpose of enabling or assisting the Board to exercise any of its functions,
- has already lawfully been made available to the public,
- is made in pursuance of an order of a court,
- is made for the purposes of a criminal investigation or criminal proceedings (whether or not in the United Kingdom),
- is made, in the interests of national security, to an Intelligence Service,
- is made with the consent of the person to whom it relates, or
- is made 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 C) This document offers the ONS Microdata Release Panel (MRP) procedure as an example of best practice for compliance with these requirements of Principle 5 (see section 5).

For non-ONS data the DPA regulates the use of personal information with authorisation for the release of private data being decided by the Data Controller. In the DPA the term ‘Data Controller’ often refers to an organisation which has ultimate responsibility for the data. Each organisation will have its own process for determining the release of data. However there are a number of ‘data protection principles’ which have to be followed. Personal data must be

- used fairly and lawfully
- used for limited, specifically stated purposes

[11] http://www.ons.gov.uk/ons/about-ons/business-transparency/freedom-of-information/what-can-i-request/approved-researcher-accreditation.html

- used in a way that is adequate, relevant and not excessive
- accurate
- kept for no longer than is absolutely necessary
- handled according to people's data protection rights
- kept safe and secure
- not transferred outside the UK without adequate protection

The DPA also defines sensitive personal data (see A3.2.5) which require greater care than other personal data.

4.2 Microdata that may be disclosive in combination with privately held data

It may be possible for confidential microdata to be disclosive when combined with additional data. These could be data obtained over time by particular individuals. Only a small number of people would have access to these additional data as they are not in the public domain so the risk is small but potentially damaging. Examples of privately held information are:

- An inquisitive individual may have built up their own database of friends, colleagues and acquaintances over the years. They could attempt to link this with published microdata.
- Personal knowledge of a friend or neighbour. Over a period of time it is easy to pick up considerable information about a wide range of people. It is conceivable that this could be used alongside a published dataset to disclose additional details.
- Response knowledge. An individual may know an acquaintance is in a dataset (they may have said that they had taken part in a particular survey for example). This could encourage a search of the data for this person with a hope that additional information could be discovered.

Microdata that are not disclosive if used with publicly available sources may be disclosive if combined with privately held data. They are neither private information under the CoP nor personal information under the SRSA, because there is minimal risk of disclosure resulting from the 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[12] for an intruder with access to 'unrestricted internet' sources (i.e. no workplace restrictions on specific sites), 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. The level of confidence the intruder has in their identification also ought to be considered. A high degree of confidence is likely to be more problematic to the data providers than an identification made with little confidence although results from Intruder testing (see Section 5.6) suggest that confidence in identifications is wide ranging and maybe the confidence level should not be regarded as an accurate indicator of a successful identification.

Protection against disclosure through combination with privately held data is provided by the restrictions of the EUL under which such data are released. The EUL is considered to provide adequate protection against disclosure resulting from the use of other data, such as private databases. A summary of the EUL terms and conditions as specified by the UK Data Service are attached as Appendix F

Microdata may be released under the EUL provided sufficient disclosure protection has been applied to ensure that they do not reveal private/personal information, taking account of the possible use of

[12]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.

data in the public domain. The Data Protection Act and the Statistics and Registration Service Act govern the use and sharing of data which contain personal, sensitive or confidential information. Data accessed through the UK Data Archive are subject to these Acts.

If required, these data may also be supplied to Eurostat. ONS also provides some datasets to customers such as Government Departments which do not want to download data from the UK Data Service under a licence giving the same protection as the EUL, for example LFS data.

Appendix A provides guidance on how to design datasets for release under EUL. Information in the appendix is also useful for EUL data not released to the UK Data Service as the guidance is consistent with the Data Protection Act as well as the Statistics and Registration Service Act.

4.3 Public use data

These data are neither private information under the CoP, nor personal information under the SRSA. The GSS microdata guidance 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. In order to protect microdata to this level, the utility is likely to be seriously reduced.

The Data Transparency agenda and advent of the Open Government Licence (OGL) could result in a significant amount of data from across the GSS being released as public use files. In many cases, they are likely to consist only of a small number of variables with more limited categories. Hence the datasets may be of little practical use for research and are likely to be used largely for teaching and training purposes. An example of a published dataset of this type is the teaching file released from Census 2011 in January 2014[13].

5. Implementation and evaluation

Sections 5.1 to 5.4 below describe the procedure for microdata release under EUL which is used by ONS. This procedure 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 section 5.2). Section 5.5 discusses briefly a framework for non-ONS releases.
Appendix A discusses the guidance procedures to follow. The terminology used in this document refers to phrases such as risk and sensitivity. These terms are introduced here to provide context.
Risk is the likelihood of an individual, business or household or other statistical unit (or related attributes) being identified in the microdata
Impact is the effect that any identification will have on the individuals, households or businesses. Data which contain sensitive variables are likely to have greater impact. Sensitive variables are those for which disclosure would cause most harm or distress to the individual concerned such as specific details about personal health or financial affairs.

5.1 ONS procedure for releasing microdata

The ONS procedure for releasing microdata is facilitated and a panel made up of senior representatives of different business areas within ONS. The panel is supported by a secretariat function. Other Government departments may have different approaches such as making data

[13] http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/census-microdata/microdata-teaching-file/index.html

available through the UK data Archive or enabling researchers to explore the data through a tabulating program. A data provider wishing to release microdata outside of ONS submits an application to the panel 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). There are 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 experts in Statistical Disclosure Control. Following a satisfactory assessment the secretariat is authorised by the panel to approve release of the data. 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 Statistical Disclosure Control experts. To help in making this assessment, the data provider is asked to complete a checklist (see Appendix E). 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 application.
Within ONS, Statistical Disclosure Control experts 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 UK Data Service 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 initial application, which is the definitive record of the data provided to the customer. The specification on this application may be used for subsequent releases of the dataset and must therefore be accurate. If additional variables or additional details are required for a future release of the data the Statistical Disclosure Control branch will need to be informed.

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

5.3 Documentation

Where disclosure control has been applied to microdata, corresponding documentation should be supplied to the UK Data Service 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 branch. 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.
Mixture of pre and post tabular processes Particular methods of applying statistical disclosure control are applicable to both tables and microdata. As highlighted in the guidance for tabular outputs from administrative data a number of Government Departments and other data suppliers allow public access to an interactive tabulation tool which can be used to generate user defined tables. It is likely that this form of dissemination will become more popular in the future allowing specific detailed tables to be extracted. Techniques have to be applied to the underlying microdata to ensure that these tables are not disclosive. Generally the microdata are protected to some degree before the tables are generated. See Appendix A3.2.7 for more detail.

5.5 GSS Procedure for releasing microdata

In practical terms there is little difference between procedures for publishing ONS microdata and non-ONS microdata. The same risk assessments should be carried out as described in Appendix 1 and the ultimate aim is to publish data of high utility and low risk.

One difference as discussed in Section 3.2 is that the SRSA is not applicable to non-ONS outputs so departments ought to find out which legislation is relevant to them. The CoP will apply in most cases with similar requirements for the protection of private information as the SRSA has for personal information.

For both ONS and non-ONS outputs the DPA gives guidance in terms of sensitive variables but lists no penalties for releasing confidential data. Differences in the Acts may lead to a (contested) assumption that the microdata release position is less restrictive for GSS (but non-ONS) data as there are clearly defined penalties under the SRSA. However from a practical point of view there will probably be little difference in the level of detail in the outputs as the public may be reluctant to respond to surveys if they suspect that the microdata are not assessed rigorously prior to release.

5.6 Intruder Testing

Prior to releasing the microdata a test to see what an intruder could find out from the data can be carried out as recommended by the ICO Anonymisation Code of Practice. This involves simulating intruder conditions (i.e. giving access to the data and relevant internet sources) to a number of volunteers, ideally with some knowledge of the data. If they can identify correctly individuals or households (and previously unknown attributes) in the data in a pre-defined time then further protection is required. The intruders are required to sign a confidentiality declaration.

This is a relatively new approach which can be applied to both census and survey microdata. Guidelines for intruder testing are being developed and will be made available when completed. For survey data one option is to give the 'intruders' a list of people who have taken part in the survey and see if they can be identified.

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 GSS/GSR 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 MRP. This panel takes advice from various relevant areas within ONS, including the SDC team, the business areas, and security and legal units.

The ONS SDC Methodology branch is able to help and offer advice to users and data providers where necessary. Their contact email is SDC.queries@ons.gsi.gov.uk.

Appendix A: 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 for ONS data releases. The DPA is to be followed for all Government releases taking particular care of the sensitive personal variables.
This appendix 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 producing microdata to be released by the UK Data Service under the 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

Anonymisation Anonymised data are suitable for publication. This requires more than the removal of the direct identifiers from the data as identification may be possible by considering the key variables. Anonymisation is sometimes defined as only the removal of the direct identifiers so it is important to be certain what is meant by this word in any given instance.

Direct identifier A variable which directly identities an individual or statistical unit. Examples are name, address, National Insurance number, NHS number.

Disclosure risk The probability of an intruder making a correct identification.

End user licence (EUL) The procedure for publishing non personal government and public sector record level data with considerable restrictions on sharing and use. EUL data are generally more detailed then Open Government Licence data

Identification Determining with a high degree of certainty that a record relates to a specific individual by using one or more variables. New information is then learnt from the remaining variables.

Imputation Replacing missing or incorrect values in the data with plausible values.

Indirect identifier Variables which allow the identity of an individual or household in the microdata to be inferred with a high degree of probability. These are usually Key variables.

Intruder Someone who deliberately or inadvertently determines confidential information about a respondent, or attempts to do so. Equivalent to an attacker.

Intruder testing People with knowledge of the relevant data attempt to identify individuals or other units under test conditions. This will give an empirical measure of the disclosure risk of the data.

Identification Key A combination of key variables

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.

Open government licence (OGL) The procedure for publishing non personal government and public sector record level data with no restrictions on sharing and use.OGL data are generally less detailed than End User Licence data.

Perturbation Changing the value of a data item either by changing the category to which it belongs or adding/subtracting a random number to the value.

Risk The likelihood of identification of a statistical unit in the data by an intruder

Population unique A record for which the identification key is unique within the underlying population.

Post Randomisation Misclassify some of the categorical variables in the data using fixed misclassification probabilities, then release the partly misclassified data together with the probabilities.

Pseudonymisation Replacement of values within identifying variables in the data with artificially-generated values

Sample unique A record for which the identification key is unique within a dataset.

Scenario A situation which allows an intruder to gain confidential information such as using public datasets or spontaneous recognition

Sensitivity Used to define variables in the microdata. Variables of greatest sensitivity are those which if released would cause the maximum distress to the individuals concerned. Variables relating to individual health could be considered to be of high sensitivity. Other potentially sensitive variables are noted in Appendix A2.

Suppression Protecting a data item by replacing the value with a symbol.

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 most likely to be used by an intruder. Certain scenarios are likely to be more common for specific datasets.

The definitions of scenarios and the corresponding key variables were developed by Angela Dale and Mark Elliot. (Elliot and Dale, 1998; Elliot and Dale, 1999) and expanded in later papers (Elliot, Mackey and Purdam; 2011). 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 (who has access to the EUL dataset) in possession of published datasets can use key variables to match these against the microdata. This may enable them to identify an individual. Examples of such published information are:

- The Electoral Register.
- Vital registration – live births, deaths, marriages, divorces
- 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

Consumer and market analysis firms such as CACI Ltd. produce a wide range out of outputs obtained from a wide range of survey and census. These are available to purchase often at geographies as low as postcode level. Examples of these datasets include consumer segmentation data and population and expenditure estimates.

One of these datasets produces modelled estimates of household income along with a range of other variables. 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 – potentially highly sensitive. Individuals belonging to an ethnic group in a particular area may feel more vulnerable if they can be highlighted in the data
- 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) – Maybe sensitive depending on the coding. A prison officer or police officer in parts of the country may not want that information to be known to the wider public.
- Income (banded) – possibly sensitive even if banded
- Qualifications – possibly sensitive. An individual might not want knowledge of their educational background in the public domain

Thus Scenario 1 assumes that an intruder could attempt to 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 is no guarantee of these correct matches although some records would be correctly matched. The

perception of a match being correct is also of importance. For variables defined as sensitive this could lead to personal details being linked to a name and address thus maximising the distress caused.

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 – in the case of high earners
- Occupation – job title, which may be equivalent to 4-digit occupation and industry codings

Other variables may be considered visible for some respondents, e.g. ethnic group, religion, accommodation type etc.

All scenarios (and especially A2.1.2) can have increased likelihood and risk if the intruder has response knowledge, that is, that s/he has some (normally private) knowledge that a given individual is a respondent and is included somewhere within the data. For example somebody who is a neighbour of a government employee could inform them that they (the neighbour) have just completed one of their surveys. This scenario can actually make smaller samples more risky as the known sample member will be more distinct in the data, though the overall disclosure risk is nearly always greater for larger sample sizes as described in A2.3).

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 identification risk comes from individuals within the microdata that are both sample uniques and population uniques on the key variables since this increases the probability of a match and therefore re-identification. A sample unique could be presumed to be a population unique by an intruder and in some cases this would be a correct assumption. There are also methodologies in place which estimate the probability of a sample unique being a population unique. The disclosure risk comes from the other variables in the EUL dataset. Provided that the identification risk is reasonably small, the data may be considered not to be personal information; this takes account of the “disproportionate effort” rule (see section 4)

It follows that the variables which are most likely to need protection to prevent identification include demographic indicators, such as geography, household composition, ethnicity, occupation, sexual identity, 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 income data are 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 senior Civil Servants, along with very high salaries and bonuses, which are all published. It could be the case that some of this information at an aggregated level could be discovered through searching the relevant websites but to link a salary to a specific individual will require access to the microdata.

A2.3 Sample size

Microdata based on a larger sample will have a greater absolute risk of identification 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 with the riskiest scenario being when a relatively large sample of a particular subgroup being taken. Microdata from these subsamples will be potentially more disclosive.

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, such as 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 for single year of age. (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 LFS where each wave takes part in four successive quarterly surveys, the Survey of Income and Living Conditions, where each wave takes part in four successive annual surveys (the Family Resources Survey is wave one), and the Wealth and Assets Survey (WAS) which is a biennial panel survey. 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 there will be a requirement for EUL microdata to permit users to link sample members across successive years. Suggestions for dealing with longitudinal datasets are given below in A3.2.6. This is a complex problem into which there is ongoing research. Any disclosure control techniques applied must enable valid conclusions to be made from the linked data.

A3. Disclosure control

Having considered the likely intruder scenarios and identified the risk factors, the process of preparing microdata which are 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.

Anonymising by removing direct identifiers is not necessarily sufficient protection. The data can still be personal information. Unique and rare combinations of variables may still be present in the data thus enabling identification of individuals/households with these characteristics.

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). Similar work is currently being carried out on microdata to be released from Census 2011. More recent work has focussed on releases from survey data for the Scottish Government and the Department of Energy and Climate Change. References will be added to this document in the near future. The advice may be updated for future microdata releases if the SDC team becomes aware of relevant published information which could be used by an intruder (see section 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 more coarsely 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 or defined subpopulation, then most key variables may not need to be protected. However there will always be some, such as geography, which need treatment, see suggestions in A3.2.3.

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. The ONS Longitudinal study is a 1% sample.

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. A subset of the sample could contain a larger percentage of a sub-population. In these cases it may be necessary to remove some records or locally protect specific key variables.

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 Region (formerly referred to as GOR). One of the main reasons for setting up the SL (section 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 (if unusual for that broader area) may reveal a lower level of geography. Care needs also to be taken with the inclusion of variables such as Council Tax and associated variables. Because local authorities publish their rates of council tax, this could reveal the local authority.

Geo-demographic segmentation classifications such as ACORN and MOSAIC, and the ONS Output Area Classification can aid identification where there are few areas of a particular category in a Region. Records with such a category should have both the category and the Region suppressed.

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 such as sub-national data for Scotland or Wales. In such cases, care will need to be taken that the level of detail of other variables is correspondingly reduced. Recoding certain highly identifiable variables (age of individual, family/household structure) will enable data to be published at a more detailed level of geography. 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 so that the original values cannot be discovered. The number of key variables in the dataset ought also to be considered when applying protection.

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 will generally be Region, (Wales or Scotland as a country). If sub national data are required (for example for Wales or Scotland) it may be necessary to reduce the level of detail in some of the other variables as discussed in A3.2.2.
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 more than 250 possible values of these	Consider whether this level of detail is needed. It may be

	variables.	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[14].	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 normally be rounded to nearest £1,000. 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 discoverable, so may help reveal a lower geography in combination with Region or other geography variable.	Large winnings should be top-coded at £500,000 (based on 2008 values). For variables based on Council Tax, see A3.2.4

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 Living Costs and Food Survey (LCFS)

(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 LCFS 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.

[14] 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.

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 Region level (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 LCFS team:

(1) In each Region, 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.

Section 2 of the Data Protection Act[15] defines a number of variables as being 'sensitive personal data'. This is personal consisting of information as to

- The racial or ethnic origin of the data subject
- His political opinions
- His religious beliefs or other beliefs of a similar nature
- Whether he is a member of a trade union
- His physical or mental health or condition
- His sexual life
- The commission or alleged commission by him of any offence
- Any proceedings for any offence committed or alleged to have been committed by him, the disposal of such proceedings or the sentence of any court in such proceedings

If these are to be published, consideration must be give to any harm or distress that may be caused. In addition a variable such as Case number could allow linkage with another dataset and any free text should be looked at carefully to see what it reveals.

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.

[15] http://www.legislation.gov.uk/ukpga/1998/29/section/2

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 Region for whom the variable has the same value.

Sexual identity

In some cases this could be regarded as the most sensitive of variables. Many people would not want this information to be public knowledge. It needs to be addressed as 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 responding 'bisexual' and 'other' 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 at all. 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 SL datasets, i.e. datasets provided to Approved Researchers, GSS standards have previously recommended that bisexual should be recoded as part of 'other'. However responses to the sexual identity question in the IHS (Integrated Household Survey) have shown the number of bisexual responses to be of the same order as homosexual responses. Additionally this group has expressed a wish for full recognition in data analyses. Accordingly it is now acceptable for bisexual to be included as a separate category in SL datasets.

Variables that reveal the sexual identity of respondents who are homosexual but are not in a civil partnership need protection. In the LFS, the livtog and livwth variables along with gender disclose same-sex couples who are not in civil partnerships, and so should not be included in EUL datasets.

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 can be described as personal details and opinions which can be included in the responses to non specific questions in a survey. Free text variables should never be included in EUL microdata and release would only be possible to Approved Researchers.

A3.2.6 Longitudinal surveys

Longitudinal surveys, or surveys with a longitudinal element, such as the LFS and the Wealth and Assets Survey (WAS), 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. For WAS the geography variable is removed from the EUL data set to ensure the variables of most interest (those related to wealth) is retained in the data in detail.

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.

The following scenario explores a possible (although possibly unusual) scenario

- Wave 1: A household has 3 residents.(A married to B; child C)
- Between Wave 1 and 2 A and B split up and D moves in
- Wave 2: The household has 3 residents (A, D and C)

If former resident B had access to the microdata most likely through being a registered user they would be able to identify the household through knowledge of the household and find out personal information about individual D. Likewise D (if a registered user) would be able to identify the household and find personal details about B from wave 1 information.

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 ONS Approved Researcher protocol or allow access through a secure environment in the relevant department.

A3.2.7 Pre tabular methods for tabular releases
It is increasingly common for users to create their own tabulations from the underlying microdata. In many of these cases the microdata have been protected by the methods described here to ensure that the resulting tables are not disclosive. For example geography may have been coded to a higher level and age banded into groups.

An alternative approach is to apply a method which has been used to produce tables from the Australian Census. This is a two stage process which initially adds perturbations to the cell values following the assignment to each record of a unique key value and then restoring additivity to the final table. This approach introduces consistency so that a cell created from the same records will always have the same perturbation added.

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 lower 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 less than 0.2% of individuals. This is not dissimilar to the 2011 Census where 0.3% of people in England and Wales live in a household of size 8 or more. 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 lower than ten. The SDC team will therefore keep this situation under review.

A5. Summary

This appendix provides guidance on preparing non-disclosive microdata, taking into account data in the public domain. ONS data may therefore be released under an EUL as being neither private information under the CoP nor personal data under the SRSA. ONS procedures require an application to release such data to include a risk assessment from the SDC team confirming that the data are not personal information. All government data (ONS and other departments) should follow the advice in the DPA with particular reference to the defined sensitive variables.

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.

The aim when producing microdata to be released under licence is that the risk of disclosure is low and the data are of high utility. Reducing risk completely is infeasible as the data would be so limited to be of little use. The defined level of risk is a subjective matter and it is not straightforward to find a suitable level of disclosure control to apply.
A post-hoc method is intruder testing as discussed in Section 5.6 which quantifies the likelihood of a motivated intruder identify a member of the sample. A number of correct identifications would suggest disclosure control had been applied too weakly. However no correct identifications could result from the correct level or too much disclosure control being applied.
It is possible to estimate the possibility of a sample unique being a population unique by determining the approximate number of people in the population with the characteristics of the sample member and thus develop formal procedures for identifying the riskiness of a record and thus the dataset. There is current research into this area and later Case Studies may include some examples.

Case Studies 1 to 3
Please see the associated document ‘Case Studies for microdata produced from social surveys’.

Appendix B: 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. (Internal ONS report)

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

Elliot, M., Mackey E., and Purdam K. (2011) Formalising the Selection of Key Variables in Disclosure Risk Scenarios. Int. Stat. Inst. Proc 58th World Statistical Congress, Dublin

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.*
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.75.8852&rep=rep1&type=pdf

Gross, B., Guiblin, P., Merrett, K. (2004) Risk Assessment of the Individual Sample of Anonymised Records (SAR) from the 2001 Census.
http://www.ccsr.ac.uk/sars/guide/2001/Gross2.pdf

Consultation on Census 2001 SAR
http://www.ons.gov.uk/ons/guide-method/census/census-2001/data-and-products/data-and-product-catalogue/microdata/samples-of-anonymised-records/consultation/index.html

Microdata Teaching File from Census 2011
http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/census-microdata/microdata-teaching-file/index.html

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

UK Data Service End-User Licence terms and conditions of access
http://ukdataservice.ac.uk/get-data/how-to-access/conditions/eul.aspx

Appendix C: 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 D: 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, (Since repealed by the Crime and Terrorism Act 2008)
(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 E: SDC microdata checklist for non disclosive data - summary

Sections 5.1 to 5.4 discuss the steps that have to be followed when non disclosive ONS microdata are published. These include a checklist to be completed by the data provider. This checklist is then assessed by the Statistical Disclosure team and this assessment is attached to the MRP application (i.e. the full application which is to be submitted to the Microdata Release Panel). The panel then decides whether to accept or reject the application.

The full questionnaire is available on request from sdc.queries@ons.gsi.gov.uk. It was felt to be too specific to be included in this general guidance document.

The introduction states

> 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.

There are specific questions relating to

- General information (name of sample, time period, level of geography, sampling design, sampling fraction, previous outputs)
- Geographical variables (Level of geography of residence and place of work, if applicable)
- Variables relating to households (size of household, type of household, number of rooms)
- Variables relating to individuals (Age, gender, ethnicity, occupation)
- Other variables
- Sensitive variables

For many questions answers can be selected from a drop down menu. For specific questions a more detailed response can be given.

Appendix F: Summary of End User Licence

Sixteen points to help you understand the End User Licence (EUL). These pointers are for general guidance and you must read and understand the full EUL before agreeing to it. By accepting the EUL, you agree:

1. to use the data in accordance with the EUL and to notify the UK Data Service of any breach you are aware of
2. not to use the data for commercial purposes without obtaining permission and, where relevant, an appropriate licence if commercial use of the data is required
3. that the EUL does not transfer any interest in intellectual property to you
4. that the EUL and data collections are provided without warranty or liability of any kind
5. to abide by any further conditions notified to you
6. to give access to the data collections only to registered users (who have accepted the terms and conditions, including any relevant further conditions). There are some exceptions relating to teaching.
7. to ensure that the means of access to the data (such as passwords) are kept secure and not disclosed to anyone else
8. to preserve the confidentiality of, and not attempt to identify, individuals, households or organisations in the data
9. to use the correct methods of citation and acknowledgement in publications
10. to send the UK Data Service bibliographic details of any published work based on our data collections
11. that personal data about you may be held for validation and statistical purposes and to manage the service, and that these data may be passed on to other parties
12. to notify the UK Data Service of any errors discovered in the data collections
13. that personal data submitted by you are accurate to the best of your knowledge and kept up to date by you
14. to meet any charges that may apply
15. to offer for deposit any new data collections which have been derived from the materials supplied
16. that any breach of the EUL will lead to immediate termination of your access to the services and could result in legal action against you