ROYAL STATISTICAL SOCIETY

Information Commissioner's Office consultation on its anonymisation code of practice

The Royal Statistical Society responded to this consultation using the document provided by the Information Commissioner. The completed document is annexed.

Information Commissioner's Office

Consultation: Anonymisation code of practice

31 May 2012 to 23 August 2012

ico.
Information Commissioner's Office

Introduction

Introducing our consultation on the ICO's draft Anonymisation code of practice

The code of practice will provide guidance on how to assess the risks of identification and how information can be successfully anonymised.

The code is intended to demonstrate that the effective anonymisation of personal data is possible, desirable and can help society to ensure the availability of rich data resources whilst protecting individuals' privacy.

Anonymisation is of particular relevance now, given the increased amount of information being made publicly available through open data initiatives and through individuals posting their own personal data online.

The purpose and scope of this consultation

Anonymisation techniques can convert personal data into a form so that individuals are no longer identifiable. The consultation will be relevant to any organisation that wants to release anonymised data, for example under the government's open data agenda.

The consultation will play an important role in making sure that the new code achieves the right balance between the protection of individuals' privacy and the benefits of making information publicly available.

The final document

The closing date of this consultation is the 23 August 2012. We are aiming to publish the finished code by September 2012.

The document will be published on the ICO website and hard copies can be made available upon request.

A summary of consultation responses will also be published on the ICO website at the same time.

How to take part in this consultation

We welcome your responses to this consultation paper.

Responses to this consultation must be submitted by 23 August 2012. You can submit your responses in one of the following ways:

Download this document and email to:
consultations@ico.gsi.gov.uk

Print off this document and post to: Data Protection Policy Delivery team, Information Commissioner's Office, Wycliffe House, Water Lane, Cheshire, SK9 5AF; **or fax a copy to** 01625 545808.

Request a copy of this document to be posted to you and post or fax it back to us. To request a copy, you can either telephone 0303 123 1113 and ask to speak to a member of the Data Protection Policy Delivery team, or email consultations@ico.gsi.gov.uk.

Please post back your completed document to Data Protection Policy Delivery team, Information Commissioner's Office, Wycliffe House, Water Lane, Cheshire, SK9 5AF. Alternatively, or you can fax a copy to 01625 545808.

If you would like further information on the code, or would like a copy of the draft code and/or consultation document in an alternative format, please telephone 0303 123 1113 and ask to speak to a member of the Data Protection Policy Delivery team, or email consultations@ico.gsi.gov.uk.

Accessibility

The final document will be published in English with accreditation from the Plain Language Commission.

The ICO has a Translations Policy that covers the publications and correspondence it produces. The policy states that, on request, the ICO will arrange for written information to be made available in Braille or on tape for blind or visually impaired users.

The ICO website also has a Browsealoud feature that reads web pages for people who find it difficult to read online.

As a publicly funded organisation we do not have the budget to undertake translation of all publications as a matter of course, but we will respond to individual requests in line with our Translations Policy, which can be found on our website.

Privacy statement

Following the end of the consultation we shall publish a paper summarising the responses. Information you provide in your response to this consultation, including personal information, may be published or disclosed in accordance with the Freedom of Information Act 2000 (FOIA). If you want the information that you provide to be treated as confidential, please tell us but be aware that, under the FOIA, we cannot guarantee confidentiality

If you are replying as an individual, the ICO will process your personal data in accordance with the Data Protection Act 1998 and this will mean that if you request confidentiality your personal information will not be disclosed to third parties.

Section 1: Your views

Please provide us with your views by answering the following questions.

1. **Do we adequately explain how the Data Protection Act relates to the issue of anonymisation?**

☐ Yes

X No
Please explain why:

The code's stated driver is the Data Protection Act, and this is clear within the document. However, given the wide ranging ambit of open data (and the acknowledgement in this code of the relevance of statistics and implicitly of statistical disclosure control), it is unclear why the code restricts its focus to just the Data Protection Act, with secondary references to Freedom of Information Acts. Some clarification of this rationale and the role of the code is needed when other important drivers could be cited. For example, the Government Statistical Service has much broader drivers, in particular the professional code of practice (and its confidentiality principle), driven particularly by ethical considerations, and other acts, including the Statistics and Registration Service Act.

2. Does the code explain adequately what anonymisation is, its technical aspects and how it's used in practice?

☐ Yes

X No

Please explain why:

The code is set out in clear, accessible language and does explain adequately what the objective of anonymisation is. The code makes a start at introducing the technical aspects of anonymisation and how it is used in practice, mainly through Appendix 1.

However, as it stands, the code is rather less successful at meeting its aim (presented at the beginning of the document) to "demonstrate that the effective anonymisation of personal data is possible". At present, much of the content of the code (for example, last paragraph on page 9 or section 3) emphasises the difficulties in ensuring that anonymisation is effective. This emphasis on the difficulty of anonymisation seems appropriate and reasonable but, without reference to more substantial examples demonstrating effective anonymisation, the code may fall short of its headline aim to demonstrate to users that anonymisation is possible (see also the society's response to Q5 and Q10).

3. Are there any key anonymisation techniques which the code does not cover?

☐ No

X Yes
Please outline:

There are 3 areas on which the code could include additional guidance, namely:

- **Sampling** - the report does not refer to sampling. In some cases, when very large numbers of records are available, it can be adequate for statistical purposes to release a sample of records, selected through some stated randomized procedure. By not releasing specific details of the sample, data holders can minimise the risk of accidental disclosure.
- **Tabular reporting** - Appendix 1 refers to methods for microdata. More could be said about ways of producing tabular (aggregated) data, which protects against disclosure.
- **Cell suppression** - A connection could be made between considerations of data quality when data are released and considerations of disclosure protection. For example, if data are from a sample survey then it would be inappropriate to release tabular outputs with cells which contain small numbers of individuals, say below 30, since the sampling error on such cell estimates would typically be too large to make the estimates useful for statistical purposes. In this case, suppression of cells with small numbers for quality purposes acts in tandem with suppression for disclosure purposes.

4. Does the code strike the right balance between the protection of individuals' privacy and the benefits of making information publicly available?

X Yes

☐ No
Please explain why:

5. Does the code cover the use of anonymisation techniques in all the key sectors?

☐ Yes

X No
Please give details:

The document would be much stronger if it could point to a number of detailed case studies in different sectors where effective anonymisation was demonstrated. At present, as noted in our response to Q2, much of the code outlines the difficulties in undertaking anonymisation, and the code could benefit from additional examples where anonymisation has been successful.

6. **Do we satisfactorily explain the issue of anonymisation and spatial information?**

☐ Yes

X No

Please explain why:

Section 5 ("personal data and spatial information") of the code is helpful in detailing the characteristics of different postcode units. However, this section would benefit from examples which illustrate the type of scenarios which follow each of the principles listed on page 27.

Specifically:

- It is not clear what is meant by the term '**statistical comfort zone**' on page 27. The paragraph that distinguishes between "statistical comfort zones" for particular geographical areas and "other forms of information that pose a risk to individuals" would benefit from examples which make this distinction clearer. At present, this section seems to encourage consideration of areas as small as 5 or 10 individuals, whereas standard examples of disclosure control practice for (admittedly highly multivariate) microdata release by government might relate more to areas no smaller than about 100,000 individuals. If the code is meant for general purpose use then it should acknowledge such variations in practice and comment on their applicability.
- The first bullet point on page 27 'The larger the number of properties or occupants in a mapping area, the lower the privacy risk' is perhaps too vague. For example, it does not draw attention to the fact that **small visible minorities in even quite large geographic areas may be identifiable**. For example if a postal district of 8,600 only includes one household with twins in their twenties or one family from an ethnic minority grouping.
- The illustration of the "**heat map**" on page 28 is useful as an example of ways in demonstrating how information can be presented in non-disclosive ways. However, as it stands, the example is not clear because it does not provide any context, e.g. what was the question? Why was this method of presentation chosen? What does the "heat map" actually mean?
- This section also states that 'Privacy risk depends on the **frequency of publishing data**', highlighting the greater

risk to privacy of “real-time” or frequently published data. However, it would ideally be helpful to have additional guidance on privacy risk relating to historic data e.g. is it permissible to provide postcode level data on where individuals lived 30 years ago e.g. precise childhood locations for cohort members now in adult life?

7. **Does the code adequately explain the difference between publication and limited forms of disclosure?**

☐ Yes

X No

Please explain why:

The Society feels that Section 7 ("Publication and limited disclosure") would **benefit from examples of application**, particularly relating to the bullet-pointed safeguards on page 32.

Additionally:

- The term '**limited disclosure**' is confusing – our reading of what is being referred to in the code is more to do with limiting *access* than limiting *disclosure*? The distinction between open data and data for which access is controlled in some way seems an important one to make.
- There is a need for some examples of **licensing arrangements and access controls**. For example, approaches adopted by the Economic and Social Data Service (ESDS) could be cited (www.esds.ac.uk). It is not obvious why there is emphasis on a 'closed community'. In an ESDS setting, each researcher accesses data from the data provider but is not permitted to share the data with a wider researcher community.
- It would be helpful to be clearer that judgements about anonymisation and methods to be used depend on the **purpose or trigger for release**. This is briefly referred to in the code (e.g. diagram on page 37), but could be pushed further. The goal of sharing for research purposes should lead to a different approach from that used with an FOIA request. In this context, section 7 of the code is extremely important and could be strengthened, to make clearer the value of this. Section 7 in fact conflates two issues around publication and limited disclosure. The last two paragraphs of the section (on the relationship between safety and utility and techniques for restricting published data) really concern publication in general and not just limited disclosure routes, but in context could be read as referring to the latter, whereas in practice limited disclosure suggests that we could continue to be able to

release data to more limited audiences with fewer restrictions imposed. The document would be clearer if the issues were separated out into different sections.

- There is also a distinction between limited disclosure which relies in part on greater trust in the **behaviour of people** receiving the data, in part because of potential sanctions for misuse, and release using technical mechanisms which control access. One of the most important features of these contexts is that it controls the availability of other information which would permit re-identification.

8. Is the section 33 research exemption clearly explained?

☐ Yes

X No

Please explain why:

The Society feels that, given the great interest currently in being able to analyse administrative microdata and to link it with survey data to create rich research resources, it would be helpful to have more explicit guidance on the issue of anonymising microdata for use in research. This gets a bit lost in the document as it is currently structured.

The suggestion in the code (page 38) that it is good practice to anonymise personal data as early as possible in the research process requires further qualification. It is important to distinguish between controlled access to personal identifiers within a research team, which is extremely important, and complete deletion of personal identifiers. Provided data have been collected from participants with appropriate consents, personal identifiers should not be deleted until all possibilities of needing to re-contact participants are exhausted. This is obvious in ongoing longitudinal studies and other clear examples would be medical studies where there is an ethical obligation to re-contact participants when analysis of samples identifies an undiagnosed condition. However there may be many other situations where further use of personal identifiers is desirable.

9. Is the flow diagram useful?

☐ Yes

X No

Please explain why:

The society believes that it is extremely helpful to have a flow diagram within the code to outline the stages required in assessing the risk of identification.

However, the existing diagram (on page 37) would benefit from added clarity, with cross references to the techniques outlined in appendix 3. An additional flow-chart outlining when each of the options in appendix 3 might be appropriate would be beneficial and may help users reach conclusions on the most appropriate form of anonymisation.

The "triggers for sharing information" box appears "out of the blue" without prior explanation and does not appear to have any obvious link to the flow chart. Importantly, it is the *question that is being asked* which is key and will often determine the route to be taken regarding anonymisation, and this could be noted in the first stage of the flow-chart. As noted under the society's response to Q3, more could be made on the production of tabular output as a means for limiting disclosure. The flow-chart could apply equally to microdata and tabular data. In the cases of tabular data, there are other potential triggers, e.g. the production of official statistical publications.

10. Do you think further diagrams, examples or case studies should be used?

X Yes

Please provide examples:

The society agrees with the aim of the code to share good practice. However, as noted in the society's response to previous questions, the code does not currently say much about where the body of good practice comes from. It would strengthen the code if it could more explicitly acknowledge those bodies of practice which are being recommended as sources of examples of good practice. References to more substantial case studies would make the document much stronger, particularly in the methods outlined in appendix 3, within section 5 and section 7. If insufficient substantial and well-documented case studies are yet available to refer to then the document could acknowledge this and note that (the relevant) guidance in the code will need updating with further experience.

As noted in the society's response to Q9, an additional flow chart to "walk through" the potential possibilities regarding the most appropriate anonymisation approach would be beneficial. At present the code outlines a range of anonymisation techniques and acknowledges the difficulties in choosing the most appropriate method, so further support to users in determining this would be useful.

☐ No

11. Is the code easy to understand?

☐ Yes

X No
Please explain how could we make it clearer:

The society believes that, in general, the code is set out in clear accessible language and does explain adequately what the objective of anonymisation is. However, there are a number of areas where suggested improvements would aid user understanding and make use of the code more successful:

- It would be helpful to include in the introduction the **definition of anonymisation** given in the glossary (page 55). This would ensure no one thinks that it meant simply removing names and addresses - the first and most basic step in ensuring that individuals cannot be identified.
- There is frequent reference in the code to the difficulty of applying anonymisation and the society feels **further use of examples** to aid understanding would be beneficial. Appendices 1 and 3 are particularly useful and could be chapters in their own right, talking users of the code through best and appropriate practice.
- The section on **borderline cases** (page 13) could be strengthened –a flow-chart outlining the anonymisation options available may help here. This section also makes reference to sensitive data. The glossary (appendix 4) provides a definition of "sensitive data" but many of the descriptions given are very broad and there is a danger that if, for example, *all* data on health is deemed sensitive this will result in unnecessarily strict controls on access to research data on health. It would be helpful to have further examples in the code demonstrating how sensitive data can be successfully anonymised.
- It would be useful to clarify in the text how the term **"pseudonymised"** is being used. It is standard practice for sample survey records (microdata) to contain a unique identifier that does not reveal the identity of the individual. This is, of course, much more important - and complex - in longitudinal studies where successive sweeps of data need to be matched to the correct individual.
- The reference to **"jigsaw attack"** on page 20 could be strengthened. As noted in the Society's response to the recent consultation on "Open Data", protection of data relating to an individual or to an individual company or organisation is a fundamental and absolute statistical principle. As ever more data is released it may become easier to piece together data from disparate sources in a way that overcomes disclosure protections. There is therefore a clear need constantly to monitor the potential for such "jigsaw" disclosure. On the other hand, it is easy to see that with datasets multiplying, concerns over "jigsaw" disclosure could become so pressing that they result in a presumption to say "no" when in fact the risk of disclosure is minimal or effectively non-

existent. Good practice will need to be developed over time and guidelines will need to be developed through regular discussion with interested parties. We suggest that one such guideline might be that individuals or companies cannot be identified indirectly without excessive cost. Under such a disclosure rule it might still be theoretically possible to identify the individual or company, but very unlikely in practice.

- All the anonymisation techniques in Appendix 1 appear to be on **microdata** and this could be made explicit in a brief introduction to this section. Also, some commentary on which techniques might be preferred in different circumstances would be helpful rather than just presenting 7 techniques. In this respect the comments at the end of each technique are helpful but these could perhaps be structured into a simple table for ease of reference. A corresponding appendix with suggestions for publishing/presenting aggregate data to ensure anonymity (e.g. heat map example) would also be helpful.
- Examples in Appendix 1 are clear but quickly become full of **statistical jargon** which is unlikely to be understood by non-statistical readers, e.g. example 7 on the "post-randomisation" method, example 8 "adding noise", and example 9 "resampling". Non-statistical audiences would struggle to understand these examples as they are current presented.
- It would be helpful if there could be a more explicit discussion in the document about the different issues to consider when publishing, displaying, or making available, **aggregate data vs.** when making **microdata** available that can be used for research and/or can be linked at an individual level with other datasets. For example the example provided on page 17 is specifically about microdata and research use whereas the heat map example on page 28 is about display of data in a way that is of intrinsic interest but could not easily be used for further research.
- The term disclosure is defined in appendix 4 and used in the code to mean the 'act of making information or data available to one or more third parties'. The society would suggest that, in this context, the term **'data release'** would be more appropriate and would avoid confusion with the term "disclosure control".
- Web-sites which do not seem to appear in Appendix 5, but are relevant, are:
    - http://www.ons.gov.uk/ons/guide-method/method-quality/quality/the-work-of-the-ons-quality-centre/risk-management/confidentiality-of-data-collected-for-statistical-purposes/index.html
    - ESDS have a number of guides on its web-site, including one on microdata handling and security: http://www.esds.ac.uk/support/datamanguides.asp

12. Is there anything else the code should cover or are there any other ways in which the code could be improved?

X Yes

Please give details:

The Society considers that a core component of the code should be that organisations making a release should consider the uses to which the data will be put and how particular anonymisation techniques would best **maximise the utility of the data**. The requirement to discuss anonymisation techniques with the data recipient is very important and does not feature prominently in the code in its current draft.

The examples in the appendix are helpful but would all result in damage to the data that might well render it of little value to research. Generally, the examples tend to be concerned with retaining the univariate distributions of variables whereas, for most research, the relationships between variables are of very high importance. There is insufficient emphasis on the link between the methods used to reduce the risk of disclosure and the purposes for which the data are needed. It would be useful to include a discussion of variables that are particularly likely to allow identification (which would include age, gender, postcode and ethnicity) and then a discussion of how to achieve anonymity in the light of the required analysis.

☐ No

Section 3: About you

1. Are you:

☐ A member of the public who has used our service?

☐ A member of the public who has not used our service?

☐ A representative of a public sector organisation?
Please specify:

☒ A representative of a private sector organisation?
Please specify:

A representative of a community, voluntary or charitable organisation, or of a trade body?
Please specify:
Royal Statistical Society www.rss.org.uk

The RSS is the UK's only professional and learned society devoted to the interests of statistics and statisticians. Founded in 1834, it is one of the world's most influential and prestigious statistical societies. It aims to promote public understanding of statistics and provide professional support to users of statistics and to statisticians.

☐ Other?
Please specify:

Thank you for completing this consultation.
We value your input.