About Anonymisation: for data about people

Anonymisation is a very valuable tool that allows data to be shared, thereby exploiting its huge social and economic value, whilst preserving privacy.

All organisations collect some information from their clients/service users as part and parcel of their organisational activities and increasingly they are sharing (at least some of) the data they collect. The information they collect on their clients/service users, for example their names, addresses, employment, financial and health status etc., is what is termed personal data. Personal data as described by the Data Protection Act (DPA, 1998) is *data that relates to living individuals who are or can be identified from the data.*

Organisations that want or need to share, disseminate or publish their data for secondary use are obliged under the DPA (1998), unless exempt, to transform the data in such a way as to render it anonymous and therefore no longer personal (for further information on exemptions see part iv of the DPA, 1998). Anonymising[1] data requires that identifiers are removed, obscured, aggregated and/or altered in some way. The term 'identifiers' is often misunderstood to simply mean *formal identifiers* such as the data subject's name, address and unique identification numbers e.g. a Social Security or National Health Service number. But, identifiers could in principle include any piece of information, or combination of pieces of information, that makes an individual unique in a dataset and as such vulnerable to re-identification.

Let us consider further this notion that any piece, or combination of pieces of information, could be identifying by using two examples. Example 1: let us consider the attributes age and marital status. At first glance these attributes are not obvious

[1] There are many techniques that can be used to anonymise data that fall into one of the following categories: data suppression, data substitution, data distortion, generalisation and aggregation. For further information on anonymisation techniques and anonymisation issues please refer to UKAN LIBRARY on our website under UKAN Resource. There are also Case Studies of good practice in anonymisation on the website again under UKAN Resource.

identifiers but let us imagine a case where one of the respondents in our dataset is a *sixteen year old widow.* Our implicit demographic knowledge tells us that this is a rare combination which means that if we were to publish this information for this respondent then she could potentially be re-identified by for example someone spontaneously recognising that the data corresponds to their friend or colleague or neighbour. Example 2: let us consider the attribute gender. Again, this attribute is not an obvious identifier but let us imagine a case where we have a dataset in which there is only one female respondent. The gender attribute then would be identifying for this female respondent.

As with any security measure anonymisation is not foolproof. Although a rare event, an anonymised dataset could potentially be de-anonymised by somebody who has sufficient auxiliary information. De-anonymisation is variously known as *data intrusion, the mosaic effect and jigsaw identification*. The idea is this: if one can bring extra information to the (released) anonymised data, then one might be able to piece together enough evidence to identify specific respondents, and/or to disclose certain attributes about specific respondents. Let us illustrate this point by returning to example 2, where our dataset contained data about only one woman. If a data intruder possessed the extra information that the woman was the shortest of all the respondents, then they would be able to re-identify the woman from the 'height' attribute even if the gender attribute has been removed from the data.

What other information might conceivably be available, and linked to an anonymised dataset, will be defined by how the data is shared. There are a range of ways in which data can be shared such as for example through secure data labs, secure remote access, licensing agreements (which restrict who has access to the data and / or how the data can be used) or openly on the internet. Of course there is a huge difference between, at one end of the spectrum, making data available to a small number of vetted individuals in a secure data lab and at the other end of the spectrum publishing that data as *open data* on the Internet. In the former case opportunities for de-anonymisation can, to a large degree, be controlled and limited simply because the data environment is being managed in terms of who can access the data and how. In the latter case there is much less opportunity to control the data environment. As such the potential for any data anywhere in the world to be used by a determined data intruder to de-anonymise data is much greater. Thus it is very important to ensure that data is shared in a way that is appropriate to the de-

anonymisation risk associated with it (see ICO *Anonymisation: Managing data protection risk Code of Practice* 2012). So for example, very sensitive and very detailed anonymised data should only be shared in a secure and controlled environment whilst non-sensitive and less detailed anonymised data can be shared in less controlled environments.

Along with anonymisation techniques and choosing an appropriate mechanism for sharing data one should be aware that there are other ways that data holders can reduce the likelihood of de-anonymisation. These include training and accreditation on issues such as data management, data storage, data security, consent, confidentiality and ethics.

It is also very important to recognise when anonymising data that the process of anonymisation may impact on the usefulness of data. For instance, in example 2, removing the gender of the respondents from the data could result in important generalisations being missed or incorrect inferences being made. So, one should be mindful that the overprotection of data is undesirable since there is little point in sharing and/or disseminating data that does not represent whatever it is that it is meant to represent.

Final comments - The Data Protection Act (1998) does not require anonymisation to be completely risk free. What it does require is that you mitigate the risk of identification until it is remote. Thus, it is the duty of care of any organisation sharing and disseminating data to ensure, to the best of its ability, that de-anonymising data will be an extremely hard problem for a data intruder.

In doing this it is essential that those sharing, disseminating and or publishing data:

(i) Use the best state-of-the-art practices of statisticians and computer scientists to assess and implement appropriate physical, technical, managerial and organisational security measures.
(ii) Minimise the risk of de-anonymisation whilst maximising the value of that data.
(iii) Establish a strategy for dealing with a de-anonymisation event if, in the rare event, one were to occur. This should include ensuring there is a clear traceable anonymisation audit trail (for further useful information please see ICO website http://www.ico.org.uk/for_organisations/data_protection/lose).

References

Data Protection Act (1998). http://www.legislation.gov.uk/ukpga/1998/29/contents

Information Commissioner's Office. (2012). *Anonymisation: managing data protection risk code of practice* http://www.ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation

Statistics and Registration Service Act (2007). http://www.legislation.gov.uk/ukpga/2007/18/contents