Response to 'Making Open Data Real: A Public Consultation'

From:

Dr Sarah Clark
Research Associate
School of Public Policy
University College London

This response is written in a personal capacity and does not necessarily represent the views of University College London.

Further information on the issues raised in this response can be found in a paper written by myself and Prof. Albert Weale, published by the Nuffield Trust and available at: http://www.nuffieldtrust.org.uk/sites/files/nuffield/information_governance_in_health_-_research_report-_aug11.pdf

Response

This consultation is to be welcomed: improving access to data generated and held by providers of public services and by government departments is essential for evidence-based policy, as well as being a way to increase the transparency of public services to citizens.

However, I have concerns around the issue of pseudonymised data, and in particular:

1. The lack of clarity about this issue in the consultation document.

2. The need for clarification of the justifications in law for use of pseudonymised data.

Background

Pseudonymised data is data which has had all personal identifiers (information such as name, address etc) removed but which has been allocated a code number which enables the data controller to link the data back to the individual via a 'key' which decodes the data.

This kind of data is essential for any research which seeks to link one dataset to another in order to assess the relationship between phenomena, for example the relationship between educational attainment and health, between socio-economic status and life expectancy, or between hospital treatment and mortality. This linking cannot be done with fully anonymised data, owing to the need for identifiers to enable matching and statistical verification.

Although it is impossible, in any other than highly exceptional circumstances, for anyone other than the data controller to identify individuals from pseudonymised data, it is still classified as personal and 'identifiable' data in law. As such, it is subject to the provisions of the Data Protection Act 1998 (DPA). By contrast, anonymised data is not subject to the DPA provisions.

There are various provisions in the DPA which are relevant. In regard to data which are regarded as 'sensitive' as well as 'personal', such as health information, Schedules 2 and 3 of the DPA offer the following justifications for use:

Schedule 2:
1) where consent has been obtained;
2) if processing is in the vital interests of the data subject;
3) if processing is "necessary for the exercise of … functions of a public nature exercised in the public interest by any person";
4) if the processing is necessary for the purposes of legitimate interests pursued by the data controller or by the third party or parties to whom the data are disclosed

Schedule 3:
1) where explicit consent has been obtained
2) processing is necessary for protecting the vital interests of the data subject or another person *and* it is either impossible or unreasonable for explicit consent to be obtained;
4) processing is necessary for medical purposes and is undertaken by a medical professional, or equivalent other professional who owes a duty of confidentiality with regard to patient information.

The issue of pseudonymised data is particularly important to public policy and social science researchers since much of their work depends on access to, and use of this kind of information.

1. The lack of clarity about pseudonymisation in the consultation document

Fundamentally, it is not clear whether the consultation document includes pseudonymised data in its scope or not.

Much of the consultation appears to refer only to anonymised data, however pseudonymised data is mentioned in the Foreword, where it is stated that further consideration will be given to the protection of personal data, "in particular the use of anonymisation and pseudonymisation techniques" (p4). It is unclear from this whether issues around these techniques are therefore within or without the scope of the consultation.

'Linked' or 'linking' data is then subsequently mentioned on pages 27 and 56. Recall that data can only be linked with other data if it is in pseudonymised form since anonymisation lacks any code by which items in one dataset can be linked with items in another dataset. Hence, it seems from these references to 'linked' or 'linking' data that the consultation document *is* concerned with pseudonymised as well as anonymised data.

On pages 17-18, reference is made to concerns around privacy and the possibility of actual identities being revealed. It is impossible to discern actual identities from fully anonymised data, since no identifiers remain - the data is fully coded and no way of decoding exists. This leads one to think once again that what is being referred to here is pseudonymised data.

Additionally, no definition of anonymised or pseudonymised data exists in the glossary of the consultation. I would suggest that many members of the public would not be fully aware of what is entailed by these terms, so definitions would have been helpful.

Recommendation

If further consideration is to be given to the issue of pseudonymised data, as is suggested in the Foreword, it is important that attention is paid to the discrete issues which are associated with it, and that clarity is established around the techniques used, the form of data involved, and the legal issues which arise.

It is regrettable that the consultation was unclear as to whether or not pseudonymised data was within its scope, and that it provided no definition.

2. The need for clarification of the justifications in law for use of pseudonymised data.

If 'open data' is to be 'made real' to public policy and social science researchers, then there is an urgent need for clarification of the justifications in law for when it is legitimate to use pseudonymised data.

Despite the range of justifications in Schedules 2 and 3 of the DPA (as outlined above) for use of 'personal' pseudonymised data, researchers have increasingly been forced to assume that they must either seek explicit consent from the individuals concerned or apply for special exemptions from the Data Protection Act to use any data which is not fully anonymised. Recall that fully anonymised data is useless for any research which seeks to establish relationships between different phenomena.

Seeking explicit consent for use of data from all the individuals in often very large datasets raises many problems: it can be expensive both in terms of financial and time resources, it can cause delays in starting research work, and it can jeopardise the validity of the outcomes of research due to factors such as consent bias and incomplete samples. Sometimes when faced with the momentous task of obtaining consent from hundreds or thousands of individuals whose data is contained in a large dataset, potentially valuable research projects are simply abandoned.

The assumption of 'consent or anonymise' has arisen because the alternative justifications in the DPA are very poorly, if at all, defined. In particular, the public interest justification outlined in Schedule 2 of the DPA, under which very much research might fall, is entirely without definition. This lack of clarity leaves researchers with no confidence about using the justification in case of incorrect interpretation and potential litigation.

The public interest provision is expanded upon in guidance to the EC Directive on Data Protection, which states that "Member States must also be authorized, when justified by grounds of important public interest, to derogate from the prohibition on processing sensitive categories of data where important reasons of public interest so justify in areas such as public health and social protection …. scientific research and government statistics" (EU Data Protection Directive 95/46/EC, Recital 34).

Recommendation

Opening up data for public policy and social science research would be a sure way of facilitating improvements in the evidence base for public policy making. However, the 'openness' of pseudonymised data is currently limited by lack of clarity in law as to its use. As a consequence, potentially valuable research which is in the public interest is prevented or at the least impeded.

As part of the policy work on ‘Making Open Data Real’ it would be highly beneficial if consideration could be given to clarifying the provisions in law which govern use of pseudonymised data. This does not require any *changes* in law, but rather expanded guidance on what is entailed by the provisions in Schedules 2 and 3 of the DPA, and most notably the public interest provision in Schedule 3.