“Making Open Data Real”: A Public Consultation

Response of the Royal Statistical Society

Introduction

The Royal Statistical Society (RSS) is the UK's only professional and learned society devoted to the interests of statistics and statisticians. It is also one of the most influential and prestigious statistical societies in the world. Founded in 1834, its early objectives were to those who aimed to gather and publish data about society, not distant from the modern objective of Open Data.

This response has been drafted by the RSS’s National Statistics Working Party.

Summary

The RSS fully supports the principle of open data and agrees that it can help deliver the benefits set out in the consultation document. However a number of key conditions are necessary for these benefits to be realised:

- Data must be accompanied by clear and succinct information about its provenance, context, purpose and reliability (metadata). Without this users will not be able to use the data effectively and are liable to misinterpret or over-interpret. Data must also be provided in a form that is clear and easy to understand and easy to re-use; this applies whether it is in an Excel or csv spreadsheet or in some more sophisticated format. Technical sophistication must not be bought at the price of clarity. It must be explicitly recognised that providing good metadata and supplying clear presentation require time and effort;
- All basic and administrative datasets should be provided free to users. The aim should also be to include, within a few years, basic administrative data currently charged for;
- Statistical datasets should explicitly include survey data from both business and households;
- Data must be properly accessible – presently users report that it can be very difficult to locate data in which they are interested. Significant efforts will need to be put into collating and curating data sets;
- Statistical confidentiality must be maintained in all data releases derived from statistical surveys.

Finally, the RSS would stress that securing any benefits depends crucially on the skills and capacities of those using the data to analyse it appropriately and interpret the results effectively. In pursuing the goals of open data government should also work to facilitate appropriate levels of statistical literacy among potential users (ultimately all UK citizens). This is an area in which the RSS is focusing effort through its getstats campaign (www.getstats.org.uk).

Detailed response

The following sets out responses to the questions for consultation.

1. Glossary of key terms

1. Do the definitions of the key terms go far enough or too far?

They provide a good basis on which to work. This is an evolving area and activity within it will help shape and refine definitions and interpretations. It is important, therefore, not to let debates on definitions to needlessly hold up progress.

2. Where a decision is being taken about whether to make a dataset open, what tests should be applied?

The presumption should be that the dataset is open. The RSS recognises that there will be exceptional reasons for a dataset to be closed and tests should be framed towards establishing good reason for a dataset to be closed. Statistical and individual confidentiality must be maintained (see Q3 under “An enhanced right to data”), but this should not be used as an excuse. Such datasets should still be released suitably anonymised.

3. If the costs to publish or release data are not judged to represent value for money, to what extent should the requestor be required to pay for public services data, and under what circumstances?

The consultation identifies the potential of open data to lead to benefits by empowering individuals, businesses and other organisations to make better decisions, and through entrepreneurship. Divining value for money may not, therefore, be a simple and straightforward task.

This suggests that the focus should be on ensuring that the costs of publication and release are minimised, and that in general no charge should be made to the users. This is increasingly possible through use of the internet and world wide web. A prerequisite for their effective use is the production of datasets in appropriate formats, and this is sensibly discussed in the consultation document. It is also important that future contracts and regulations are drawn up with open data in mind as problems arising from contractual, legal and regulatory issues can add significantly to official time (and hence cost) in making data public.

4. How do we get the right balance in relation to the range of organisations (providers of public services) our policy proposals apply to? What threshold would be appropriate to determine the range of public services in scope and what key criteria should inform this?

The balance will, to some extent, depend on the definition of what organisations are considered to be providers of public services. Policy proposals are best implemented when there are no grey areas, i.e. all organisations are in the range unless there is reason for them to be excluded.

Ultimately, this is an area that will be demand driven and any policy proposals will need to allow for responsiveness and adaptability. In the early days, producers and suppliers of data will already have some understanding of the level of current demand for each of their particular outputs which may allow them to focus their efforts.

5. What would be appropriate mechanisms to encourage or ensure publication of data by public service providers?

The role played by ministers will be crucial, giving instruction or applying pressure accordingly. Parliamentary or other scrutiny will provide a means of identifying those providers who are or are not responding effectively to requirements. Where the response is not sufficient then means of sanction will need to be in place, with the Information Commissioner (or others) being provided with appropriate powers.

It is also important that future legal contracts and regulations are drawn up in ways that facilitate open data subject to confidentiality requirements (see Q3 above).

An Enhanced Right to Data

1. How would we establish a stronger presumption in favour of publication than that which currently exists?

In addition to the measures taken to encourage and ensure publication of data (as set out in response to question 5, *Glossary of key terms*) it will be important to publicise examples of good practice, particularly where they have led to identifiable benefits. The aim must be to establish a culture where the presumption is ingrained within the organisation.

2. Is providing an independent body, such as the Information Commissioner, with enhanced powers and scope the most effective option for safeguarding a right to access and a right to data?

Users of data will need to know that their rights are meaningful and that they have easy means by which to have any grievances addressed by a body with appropriate powers to take actions. Independence is crucial. The arguments here are similar to those made in, and supported by, Parliament in establishing the UK Statistics Authority under the Statistics and Registration Service Act 2007.

3. Are existing safeguards to protect personal data and privacy measures adequate to regulate the Open Data agenda?

Currently these measures are adequate. Indeed some users feel that some controls currently being applied to statistical datasets are overly protectionist and result in data being denied where there is no real privacy issue. This is not an easy issue to resolve.

On the one hand, effective data collection can depend greatly on confidence in the protection of personal 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.

It may also help to recognise that some information is more sensitive than others. An individual's health information, for example, is clearly highly sensitive and needs to be protected very carefully. At the other extreme, prices charged by supermarkets and other retailers for goods are clearly already in the public domain and confidentiality concerns really arise only from the contracts through which data are collected.

It is important that whatever techniques are in place that they are applied consistently and that knowledge is shared across government departments and their agencies to ensure this is expedited. There is a role for the Office for National Statistics to give leadership in this area given its experience.

The Cabinet Office may like to consider the judgement of the House of Lords on the case of Common Services Agency vs Scottish Information Commissioner. http://www.publications.parliament.uk/pa/ld200708/ldjudgmt/jd080709/comm.pdf

This ruling identifies the issues of handling personal health data and the risks of disclosure of childhood leukaemia data.

It is, of course, important to distinguish between statistical datasets collected for statistical purposes (mainly surveys) and administrative datasets, which can be used for statistical purposes, but whose origin is in departmental administration and to which different confidentiality rules might apply. The principle ought to be consistency with the commitments to the respondents/suppliers of information. Where confidentiality has been promised then statisticians have expertise to offer.

4. What might the resource implications of an enhanced right to data be for those bodies within its scope? How do we ensure that any additional burden is proportionate to this aim?

In general, it is more costly to keep a dataset secure than to make it open, so in this respect open data could save money. In the shorter term there could be costs in some departments depending on what state the data currently exists in. An approach may be to ensure that all new data produced is in appropriate formats, with the re-formatting, where needed, of past data dealt with on a prioritised basis. Prioritisation based on user demand would provide a basis for proportionality.

It is also clear that in some cases the costs of making data "open" arise not so much from the technical problems but from legal, regulatory and contractual issues; hence the importance of ensuring that future contracts and regulations are appropriately worded to facilitate open data.

5. How will we ensure that Open Data standards are embedded in new ICT contracts?

Those involved in drawing up and agreeing contracts should be fully aware of the need and this may require specific training programmes. A process of certifying that contracts meet policy requirements might be appropriate. Clarity will be required over the standards for which central guidance may be necessary.

Setting Open Data standards

1. What is the best way to achieve compliance on high and common standards to allow usability and interoperability?

Standards developed with all stakeholders having an opportunity to inform their development have a greater chance of compliance. Stakeholders will understand why any particular standard has been laid down.

Ensuring standards are set out in a single document will maximise ease of compliance and minimise the potential for any particular one being overlooked. A code based on the Public Data Principles would be valuable. The adoption of the UK Statistics Authority's Code of Practice for Statistics may be a helpful model.

The Government Statistical Service and the UK Statistics Authority should play a leading role in spreading good practice and ensuring a consistent approach. The outputs of the GSS Transparency Sub Group should be used to this effect.

2. Is there a role for government to establish consistent standards for collecting user experience across public services?

It is not clear if this refers to user experiences of accessing data published by different public service providers, or to the collection by different public service providers of user experiences of their services. However, in both cases the exercises will be most useful if comparison can be easily and unambiguously made, and this would be facilitated by using standard methods wherever possible.

Whatever the role played through consistency of standards, other forms of user engagement (such as user groups) will remain important especially as they can provide deeper insight or help fill out an incomplete picture. Above all, standards should be framed as to be enabling, not restrictive.

3. Should we consider a scheme for accreditation of information intermediaries, and if so how might that best work?

Basic datasets should be released directly without the use of intermediaries.

It can be expected that with the emergence of an increasing number of customers who require sophisticated analysis of open data then the information intermediary sector will grow to meet that demand. Customers will want to know that the service they procure is provided by a reputable company. A body or bodies may emerge to represent companies within the sector, as with other established sectors. Membership of such bodies, particularly if they adopt standards for membership, may emerge as a form of accreditation on which customers can form an opinion. Government may have a role here in stimulating the formation of such bodies rather than directly setting up an accreditation scheme. In the earlier days, government's role may be to issue guidelines and advice, short of an accreditation scheme.

A risk with an accreditation scheme established centrally and early on is that it may divert activity from focusing on releasing data. Indeed "hands off" should be the default policy for government in this area.

Corporate and personal responsibility

1. How would we ensure that public service providers in their day to day decision-making honour a commitment to Open Data, while respecting privacy and security considerations.

Honouring commitments to Open Data on a day to day basis will be best met in which the culture of the organisation is one of doing so, and where there is a means of measuring it.

It will be vital that top-level management both require and champion these commitments and that those on lower levels both recognise their responsibilities and are rewarded for delivering on them. This may require training.

Measuring adherence to the commitment will be harder though. This may be less hard among those public service providers which already have some form of scrutiny, particularly where this is done by external body.

2. What could personal responsibility at Board-level do to ensure the right to data is being met include? Should the same person be responsible for ensuring that personal data is properly protected and that privacy issues are met?

The Caldicott approach of having one senior person responsibility for both confidentiality and data sharing sounds much better than split (and conflicting) responsibility.

3. Would we need to have a sanctions framework to enforce a right to data?

Given a right exists then a sanctions framework is a necessary requirement in support of that right. Those with grievances will need to know that these are addressed in ways that lead to outcomes that individually are timely and satisfactory, and generally lead to changed behaviours.

4. What other sectors would benefit from having a dedicated Sector Transparency Board?

This could be a promising development but must be driven by users' needs, rather than those of supplier organisations. Broad categories that have been suggested include Land and Property (addresses, maps, etc), and Social Statistics (Census, social change, etc).

Meaningful Open Data

1. How should public services make use of data inventories? What is the optimal way to develop and operate this?

Appropriately presented, inventories will be useful in providing users with a means of understanding the data available. Poorly designed they will inhibit users (particularly those who are occasional or new users) who may face overwhelming and complex data. It is important therefore that inventories are designed to help users get started. Data inventories take a great deal of resource to produce, but there is currently little evidence for their use. Understanding this lack of use will be important in refining existing inventories and developing new ones.

Furthermore, there should be a joined up approach to inventories. They are currently disparate and confusing to the user. The recent review of government data could be a start point, but this has not been published.

2. How should data be prioritised for inclusion in an inventory? How is value to be established?

Ideally an inventory is all inclusive and so any issue of prioritisation does not arise. It may be that data assets should be prioritised for their completeness of their description within the inventory. Prioritisation could take place on grounds of user demand, on potential for economic or social impact, or for supporting public scrutiny. User demand may already reflect the potentials in the latter two areas. Government might decide to do this initially to help meet its other policy objectives.

Although imperfect there will be some understanding of existing user demand among government departments and public service providers. However, as there is anticipation that open data will lead to wider and greater data use, any approach must be flexible to adapt to emerging demands.

On a specific level, parliamentary questions and freedom of information and other requests could be used to further gauge demand.

3. In what areas would you expect government to collect and publish data routinely?

If the objectives of Open Data are realised then all data should be published routinely together with appropriate metadata, except those excluded for publication for reasons based on appropriate tests or for confidentiality requirements.

Data collected should be a matter of systematic review on tests of how it meets the economic and social needs of the UK both for government and outside of government. The objective must be of serving the public good (as set out in section 7 of the Statistics and Registration Service Act).

4. What data is collected "unnecessarily"? How should these datasets be identified? Should collection be stopped?

This is not an easy question to answer, as the experience of the GSS shows. In implementing data publication, systems should be put in place to monitor its access (eg web downloads) and, where possible, secure feedback on use and usefulness. A start point would be to consider the use of data.gov.uk.

5. Should the data that government releases always be of high quality? How do we define quality? To what extent should public service providers "polish" the data they publish, if at all?

The question of definition of quality is an important one. It covers different things, including whether the data definition is appropriate for proposed use, whether coverage is sufficiently comprehensive, extent of errors, and whether format is right for ease of use. A particular dataset may have more than one potential use – it may be ideal for some uses but also useful with reservations for others. Hence there is a need for appropriate metadata. Provision of metadata, and upgrading formats, is where effort should be directed.

Very poor quality data can be meaningless and misleading, and there should be basic checks for errors. However, if the data is fit for purpose, it should be released. A simple test may be that if the data is used within government then it should be released for use outside of government.

Providing metadata that is helpful, succinct and readable is not a trivial task and one that will probably need some training. It is not a skill that comes automatically to many civil servants but the quality of the metadata will be crucial in facilitating effective use of data. It will also be important to have a consistent approach to this and to ensure that standards are adhered to. The GSS will have an important part to play.

The focus on quality should be ensuring that the data is easy to access and use. It is important to recognise that usefulness is related to timeliness, especially if competing datasets are released before those originating through government. As long as the metadata is appropriate there should be a presumption towards timely publication.

Government sets the example

1. How should government approach the release of existing data for policy and research purposes: should this be held in a central portal or held on departmental portals?

There needs to be clarity over data portals versus data storage. Data can be held in any number of places, and the portal simply provides a way to access it. A central portal is likely the best solution, carrying the following advantages:

- it would provide a single point of access, avoiding confusion over where to access data (particularly if the data production shifts between government departments and other public service providers);
- it would result in a consistent user experience;
- it would be permanent and high profile.

However, it is vital that data are accompanied by good metadata. There are challenges in ensuring this happens across government, which may involve setting standards centrally and ensuring they are implemented. This also applies to adherence to Open Data standards.

The portal must have a good search engine, and the data must be appropriately indexed to ensure they can be found by other search engines.

However, it must be ensured that development of a central portal does not hinder release of data. The experiences of users of data will play a crucial role in determining how best to implement a portal or portals and how to ensure they develop appropriately. Such experience can only come if data is available for use.

2. What factors should inform prioritisation of datasets for publication, at national, local or sector level?

This is fundamentally informed by extent of (potential) use, and hence value, and also having data at the lowest possible level (e.g. Output Areas for statistics in most

cases, or even postcode for non-sensitive data). Much potential value is destroyed if information is only available at local authority level or above.

3. Which is more important: for government to prioritise publishing a broader set of data, or existing data at a more detailed level?

One should not be prioritised over the other. Evidence of potential user need coupled with the ease of making data available should be the initial criteria. Charging for such data undermines the principle of "open to all" and should be avoided wherever possible.

Innovation with Open Data

1. Is there a role for government to stimulate innovation in the use of Open Data? If so, what is the best way to achieve this?

The consultation document sets this out succinctly: *"The best way to tap into the UK's tradition of creativity and invention is to give that data away".* Again, it is vital that data is given away with appropriate metadata.

A single, easily accessed starting point for information about government data is very important. It may help guide users to other sites where they can access the data. It may also allow for resources to be published explaining what generally might be learned from data and might provide a place for encouraging statistical literacy among non-expert users. Generally, data.gov.uk is a good start but needs further development.

The focus should be on removing the barriers to the use of data (e.g. IT constraints, charging/licensing, ensuring efficient disclosure control procedures), improving awareness of what data is available, and provision of appropriate metadata. This will encourage innovation in the private sector.

Contact details
Royal Statistical Society
12 Errol Street
London EC1Y 8LX

T: 020 7638 8998
E: rss@rss.org.uk
W: www.rss.org.uk