RESPONSE TO "MAKING OPEN DATA REAL: A PUBLIC CONSULTATION"

Michael Clary
Retired Government Statistician

I am a private individual, having retired from the Government Statistical Service just over a year ago. This means that I had considerable experience during my working life of the difficulties that can arise using data, and in particular in the use of administrative data for statistical and other analytical purposes. I also had some involvement in the immediate impact of the initial Open Data initiative on data providers, particularly in the statistical field, though of course the landscape may have changed since then.
Before attempting to answer the specific questions posed in the paper, I would like to make some general comments.

General emphasis of paper

When it was first unveiled, the Open Data initiative had several strands, each of which appeared to me to be given roughly equal importance.

- Access to public sector spending (accountability).
- Access to public service delivery information, in order to provide more informed choice, and to drive up performance (transparency).
- Access to more general datasets held by the public sector, to enable business and private individuals to develop novel applications and/or provide useful public-facing applications. The aim could be to increase economic growth, or simply improve the quality of life (development). The opening up of some Ordnance Survey data is an example.
- Enabling datasets from different sources (not necessarily all public sector) to be linked together, enabling much richer understanding of the meaning of the data (analysis).

While all four are still covered, it appears to me that the consultation paper is heavily focused on the first and second of these bullet points. There is not much in the main paper on the advantages of access to wider public sector data (though there is more in Annex 1) and data linking is hardly mentioned. This is disappointing, as it seems to me that the benefits from these areas could be considerable.

On the face of it, the change of emphasis appears to reflect the political priorities of the new government. The front page of the website reinforces this impression, with a heavy concentration on the workings of government. Fair enough, it might be argued, but I believe that there is a discussion to be had about whether an initiative of this nature should be taken forward in a more visibly independent manner. Otherwise there will always be a suspicion that datasets useful to the current government (of whatever colour) will be prioritised, while those less palatable will be promoted less, and perhaps dropped altogether. To put it another way, my right to data becomes less valuable if the source is liable to be abolished. With my statistical background, I wonder whether there are lessons to be learned from the Statistics Commission, at least in ensuring that proper process is followed when considering changes to source availability.

Meaningful open data

This is a sub-section within Section 8 of the paper but the discussion there is largely limited to providing users with a clear understanding of what exists. That is important but it seems to me that there are much more important issues which should fall under this heading, but which appear not to be touched upon in the consultation document.

In order for data to be meaningful, they need to generate reliable comparisons. The most obvious ones are comparisons over time and across geographies, but there are others – across industries and across subgroups of the population. My right to data becomes less valuable if I cannot draw worthwhile conclusions from it.

Suppose we have a nationally compiled dataset, obtained from an administrative or delivery process. One would hope that it would yield satisfactory comparisons across areas, industries, or subgroups of the population. But will it yield proper comparisons with other years. The answer will often be that it cannot, because the process has been changed regularly over time. These changes are often perfectly justifiable, and may be essential reactions to problems highlighted by earlier tranches of data. But they are likely to destroy comparability.

There are two well known examples. Unemployment data from the benefits system became notoriously difficult to compare as the system changed repeatedly. Recorded crime data are impacted by changes in guidance on how to record crime, changes in the performance of police forces in following this guidance (and of course the likelihood that a crime may be reported, which could reflect insurance-related issues).

The general (though not universal) view is that these shortcomings render the administrative data unsuitable for analysis of trends over time, and that the statistical sources available (the Labour Force Survey and the British Crime Survey) are far better for this purpose.

However, this gives rise to another problem. There is, unsurprisingly, a lot of interest in seeing data for very small areas. The statistical surveys cannot meet this requirement, as the sample sizes required would be unpopular and prohibitively expensive. So the administrative sources have to meet this need, but comparisons between small area data for different periods are unlikely to be reliable if systems have changed.

There is of course a wider issue with data based on very small numbers. The numbers of cases can fluctuate a lot from year to year purely as a result of statistical chance but users may see them as meaningful and requiring an explanation from the service deliverer.

Issues such as these can be flagged up in metadata with the original release, but experience suggests that these metadata tend to disappear from subsequent applications of the data! This is not an argument for opposing release, but there are issues which will take some getting used to, and in the worst case scenario generate large numbers of requests to explain what does not really require explanation.

Perhaps comparability will be easier to achieve if data experts such as statisticians are involved in the outset in determining what changes should be made, and how they should be implemented. There could be other advantages in involving analysts. I have seen too many examples of analysts being brought in – too late – to analyse datasets obtained without any thought that they might later be used in analysis. An apparently trivial example would be failure to record a postcode, which probably had no implications for the process itself, but could cause problems subsequently if geographical analysis is required, or if there is a need to link the data to some other source.

As a final point on nationally generated datasets, there will be inevitably be suspicion that changes of the nature described above are politically driven, in order to massage the figures. There is a role for independent oversight of such changes, in order to reassure the public (and government itself) that they are indeed soundly based.

If we move away from national to locally generated datasets, the position could be still worse. The localism agenda appears to imply that each local authority should decide (informed by residents' views) what data it should collect, compile and release. So my right to data may give me information for Enfield, but will it give me comparable data (or indeed *any* data) for Barnet and Haringey? There seems to be a tension here between the understandable desire to get away from the excesses of the national indicators for local authorities and the need (in my view) for meaningful comparisons between areas.

Two stars may do

The rating system in section 8.9 carries an implication that the highest star rating is the best. However, for many prospective users, simple PDF or Excel files may meet their needs. I would go further – if data are only available in more sophisticated formats, they may become inaccessible to much of the population. I would argue that the simpler forms of presentation should continue to be available where practicable. Of course this would not be possible for huge databases such as COINS.

I would also like some attention to be given to barriers to the use of these simpler forms. Here are some examples from the transport field.

- Some public sector sites (e g Traveline) do not use fixed urls. This prevents users bookmarking a file rather than having to plod their way through the interface provided each time. In my view this practice is only acceptable where the output is being generated on-the-fly, but this should not be the case for a complete timetable for a bus route.
- Some public sector sites insist on the user registering in order to access the data. The TfL timetables data on the London Datastore is an example of this. To quote from a blog on the Datastore site:-

  "Apologies but this data is no longer provided on the Datastore and is only available via the TfL Developer area. We recognise that this requires you to register on the site but this is current TfL policy."

  TfL have done excellent work in the open data area, but this policy seems out of kilter with the spirit of Open Data. I can see a justification for requiring

registration for anyone intending to access the data source programmatically, not least because of the potential traffic generated, but it should not be necessary simply to download the files. One of the principles of open data should be that what the end-user does with their data is none of the provider's business, unless there are implications for the performance and stability of the data provider's website.

I will now move on to the specific questions asked.

Glossary

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

The words "factual", "data" and "structured" are used on the definition of "dataset" but are not defined. Does "structured" mean that each of the data items is in a standardised format, and "unstructured" that they could be free text? Or does "structured" mean that the data are stored in (say) a spreadsheet or database? Is user satisfaction data regarded as "factual"? Is there a distinction to be drawn between raw data on individual's interactions with a provider, say, and aggregated data; the three examples given seem more likely to be the latter.

The word "information" is used in the definition of "dataset" but with a quite different meaning to that given in the definition of "information" itself!

The model I would suggest is that "data" on their own are the raw material which analysis helps turn into "information". In those terms, what is defined here as "information" might better be labelled as "analysis".

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

The obvious test is one of harm to the national interest or to individuals or businesses. Although a value-for-money test would seem sensible, in some cases it is only after the dataset has been in the wild for a number of years that its value becomes apparent. Trying to value a dataset in advance is likely to result in too restrictive an approach. I would however accept that there are some datasets where it is next to impossible to imagine a worthwhile use.

On the other hand, it would be nice if some of those proposing access to datasets could manage more than a one line explanation of the benefits!

There is also an issue as to who should make these decisions. Should it be left to individual organisations? Should there be a group within government driving forward a common approach? Should there be ministerial input into this, or should any responsible body have the same sort of detachment as (say) the Statistics Commission?

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?

Leaving aside the difficulties of making this assessment, and of doing so consistently (see #2 above), there is a difficulty here. What happens if second and third data requestors apply after the first data requestor has paid up? Do they have to pay the same or do they get the data for free because the costs have already been recovered?

The main costs are likely to be turning data that are fit for internal purpose (but were never intended to go further) into something which can be made available widely, and the costs of ensuring that nothing is revealed, directly or by implication, about individual persons or businesses. Arguably, any government that is serious about this agenda should be prepared to bear the first of these costs as a one-off. The second cost is trickier as it will reoccur every time a dataset is updated.

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

Be realistic about the resource need. Governments of all hues tend to think that their own projects can be done at zero cost, while its opponents' projects were an exorbitant waste of resource! In particular, the number of data publishers within government who have any experience or training in RDF (or anything beyond spreadsheets and PDFs) is – or at any rate a year ago was – small. Training them will be a significant cost.

The other encouragement would be to move some of the focus of the site back towards data analysis in its own right rather than as a means of enabling "armchair auditors" to highlight the more absurd spending items. While these data should certainly be available openly, the overemphasis on this aspect does sometimes give the impression that the very people expected to deliver on this are finding themselves under siege as a result.

The Executive Summary is a good example, saying next to nothing about the potential for the private sector to add value by exploiting open data. Even where it is mentioned (e g paras 4.3 and 7.8) the emphasis is on exploitation of information about public services. This is too narrow; one of the key early gains for users was the freeing up of considerable quantities of Ordnance Survey data, which most people would not think of as public service information.

Policy Challenge questions - An enhanced right to data

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

Presumably the word "publication" (I would prefer "release") is meant to imply the making available of raw datasets as well as what might be termed the results. Simply making available the tables and analyses that government has chosen to produce would greatly reduce the scope for innovation by others.

I would be very reluctant to see cost used as a reason for refusing release. I would draw a distinction between the one-off costs of changing IT systems to generate suitable formats and the ongoing costs associated with ensuring confidentiality of

personal data. The former should be borne by any government that seriously believes in this agenda. The latter are genuinely more difficult as it is hard to reduce the series to a process of mechanical steps; some human resource input is probably required on an ongoing basis.

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?

Some form of independent body is essential. Not only to ensure that data holders are not seeking to circumvent the open data regime, but also to minimise suspicion that the government of the day is holding the door open for data that suit its purposes and obstructing data which do not.

Some features of such a body could be drawn from the Statistics Commission, and it is encouraging to see (para 8.11) the parallels being drawn between open data and government statistics. It is important that such a body should not be dominated by data producers and/or by the analytical community. Open data do pose challenges in terms of personal privacy and it is important that the public does not feel that their interests are being disregarded by one or other of the special interest groups. At a severely practical level, if people have doubts about how their data are being handled, they are less likely to participate in surveys or to provide entirely frank data when interacting with service delivery.

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

I am not sure. However, I would point out that the regime in place for statistical data is rigorous, probably more rigorous than most outside the statistical arena would imagine. Data for a single individual or business must not be releases, but neither should data for multiple individuals or businesses which could reveal something to one of those "units" about another. I have doubts that anonymisation, one technique identified in the paper, is sufficient, particularly if carried out by those without experience in this tricky field. Government Statistical Service expertise should be invaluable in determining the approach to be adopted.

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?

The implications could be serious for any body which does not currently collate the potentially in scope information into a consistent format and for those faced with the ongoing burden of checking for confidentiality. I assume that this is a cost-benefit question. However, the benefit is not easy to assess in advance of making the dataset available. The private sector might manage to use the dataset in quite unpredictable applications which have huge benefits to the economy.

As an example of benefits which were probably unpredictable, at least in terms of their scale, consider the humble postcode, basically an operational convenience for

the Post Office. I wonder whether anyone envisaged just how prominent a role postcodes would come to play in mapping, geographical analysis and business applications.

Policy Challenge questions – Setting open data standards

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

I would query whether *every* dataset needs to make the transition from one (or perhaps two) star to higher ratings. Is there any benefit in making the BIS list of ministers (if that is still there!) machine readable, in open format, or linkable? It may also be that a spreadsheet is fine for most users of a statistical output, say, because they are in turn going to use a spreadsheet product to perform further analysis. Just providing RDF, irrespective of benefits (unpredictable though they may be) could waste a lot of resource.

I would also argue that simpler formats should be retained even if higher rates formats are provided, as the latter may be hard for many to use.

Finally, I note that the availability of good metadata appears not to contribute to the star ratings. It should.

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

This is put too narrowly. There should be a role for the government (or an independent body) to establish consistent standards and approaches to collecting a much wider range of data. What use are data for my borough to me if they are not consistent with the data for its neighbours, or even with its own data for previous years? Delegation of decision-making is superficially attractive, but cuts across the need to obtain maximum benefit from the data that are collected. There is a range of issues around consistency of data, which the consultation paper does not appear to touch upon.

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

This question appears unrelated to the previous discussion and I do not understand it. I would not want to see access to open data restricted to a charmed circle of practitioners. The general public should have the same rights.

Policy Challenge questions – Corporate and Personal Responsibility

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

Hopefully fear of exposure would act as a strong distinctive to obstruction by staff. However, failure in these areas is often for more complex reasons. The most obvious one is a failure to accept the resource implications. That could be a collective failing

amongst a body's staff but it could also stem from reluctance at ministerial level to make the funding follow the priorities. Perhaps sanctions against ministers should also be available!

Policy Challenge questions – Meaningful Open Data

I am disappointed to find that this section deals primarily with meaningful data *inventories*. From its title I had expected to find some sort of discussion about how to ensure that data obtained from public service delivery systems are meaningful.

My definition of "meaningful" would be wide. Data should be consistent over time. They should also be consistent over geographies. The latter is simpler to discuss. It should be relatively easy to ensure that data from a central government system are on the same basis for each local authority area, for example. It is harder to ensure that the same is true for data from each local authority's system are comparable. Have the same definitions and criteria been used? Have the data items been collected in the same way (even asking the same questions in a different order can produce different responses)?

I cover this in more detail in my general comments at the start of this response.

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

It would be wrong for a data provider to actively prioritise data for inclusion in an inventory, as this assumes they are well placed to assess the value of the data. Far better to maximise the content of the inventory and then let the market consider what may or may not be of value. I would accept that there are some datasets which have little conceivable value as open data, so there would be exceptions to my principle.

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

Clearly it may be necessary, even desirable, to stop collection where there is little prospect of it ever becoming relevant to policy or to the wider public, or where financial circumstances dictate. It is however important that any government considers fully and fairly the likely needs of a successor government and does not simply drop sources because of ideological doubts about its basis. The worst possible outcome would be a see-saw effect where what one government does is continually reversed by its successor. Much of the value of data, particularly for investigation of economic and statistical relationships, lies in the availability of a long time series of data.

I would suggest some degree of light touch independent oversight of this process, in the same way that the Statistics Commission can comment on proposed key changes to data sources (as well as on breaches of consultative procedures).

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?*

No, but quality issues must be clearly flagged up in metadata. It is worth noting that a release may be fine for large areas, or large groupings of industries, but flaky for small areas or for detailed industries.

Policy Challenge questions – 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?*

This seems to me to be a question predicated on old technology. It should not matter to the user where data are held, as long as there are adequate routes to the data from anywhere the prospective user might expect to look. For government statistics (where pre-release restrictions make it difficult to hold them anywhere other than on departmental websites) these portals would include data.gov.uk, the National Statistics publication hub and the departmental website itself.

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

I would give priority to detail, as it provides the building blocks from which the user can build up the analysis that they wish to see. I would apply this principle to data organised on both geographical and industry bases, for example, though I have noted with regret that ONS publishes less industry detail than it used to.

Policy Challenge questions – 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?*

Yes, both within the public sector and in partnership with the private sector. It will be important to avoid giving the impression that the public sector is considered as part of the problem rather than of the solution.

Annex 1

Sorry, but I cannot resist asking when we are likely to see published productivity figures for the Cabinet Office!