TONY HIRST RESPONSE (VIA EMAIL)

The following is a *personal response* to the Making Open Data Real Consultation

Tony Hirst

Open Data Consultation [http://www.cabinetoffice.gov.uk/sites/default/files/resources/Open-Data-Consultation.pdf]

Notes

***1. Do the definitions of the key terms go far enough or too far?
"Public services are either provided by public bodies, or providers who have been funded, commissioned or established by statute to provide a service"

I assume the definitions of open data and public services are to be taken together, with the consultation focussing on 'open (public) data produced by public services'? For such bodies, I assume there is also a formal "data burden" that defines the public data reporting requirements to the centre, as well as devolved data burdens eg into local government from schools? Would it make sense to clarify the notion and extent of data burdens, and the extent to which elements of these (and the organisations they apply to) should be subject to open data requirements? I guess there is also a data burden placed on individual citizens in respect of filing tax forms, for example, that are not subject to openness requirements?

A clear statement at least of data burdens/formal reporting requirements between public bodies that are in scope for mandatory release as open public data should be made available, eg along the lines
of http://www.communities.gov.uk/localgovernment/decentralisation/tacklingburdens/singledatalist/ http://getthedata.org/questions/500/data-burden-on-uk-higher-education (I know some work has already been done on this that I used as the basis for a simple data brden visualisation exercise (http://www.flickr.com/photos/psychemedia/5536836259/).)

"Dataset"
It may be useful to distinguish between data collected for operational, administrative or statistical use, as well as the extent to which data produced in the normal course of events is being legitimately requested as is, or whether it must be processed before release (eg http://www.adls.ac.uk/what-is-administrative-data-and-why-use-it-for-research/ http://www.unsiap.or.jp/ms/ms7/DennisP1_OppoChalle.pdf)

It may also be worth distinguishing between the release of complete data sets, views over the data that represent a query on a complete dataset, and queries, sampling procedures or any other means that are used to generate those data views. For example, providing data relating to performance indicators for a particular school in response to an FOI request from a citizen equates to the provision of a particular view over the database containing perfomance indicators for UK schools as a whole; providing a copy of the database as a whole to a developer of a school comparison website represents the provision of a complete dataset.

Datasets may provide value to others in a variety of ways: for example:

- using complete datasets as the basis of comparison or recommendation services;
- using complete datasets to support statistical analyses, segmentation/clustering of data;
- generating very particular or specific views over the dataset by constructing meaningful and appropriate queries on the datasets. Queries are also reusable, and whilst some cost may be incurred in creating them, making them open, and suitably parameterising them, the marginal cost of reusing the queries is then minimal. It is possible that queries that take a long time to create/optimise become valuable in their own right, and that the dataset and the view can be given away freely. The query unlocks value in the dataset and delivers it to the requester. When it comes to government reporting, where reports include summary views over open datasets, the openness/transparency requirement should not deem to be met unless the query that generates the view from the dataset is also openly published.

Datasets may also include recordsets relating to an individual; where personal access to personal data/mydata is possible, we need to distinguish between the private/personal right for an individual to access their data, or an agent acting on their behalf and with their permission, as opposed to general public access.

Where public monies are used to find data acquisition, arguments can be made in favour of that data being mae public as a consequence. As far as UK Research Council funding of academic research goes, there is an increasing requirement for open access publication of research results. However, there are not necessarily any similar requirements around the opening of research data. It is typically the case that "negative" research results are not published, and as a consequence the data collected is also unlikely to be made available. If the costs associated with archiving that data in an open and public way are marginal, or are covered as a funding requirement whether or not a research project is "successful", then *all* research data could be made public as long as the method is sound, irrespective of whether "significant" findings are discovered relating to the particular research question asked.

Where public body activities are covered by the Audit Commission, presumably an argument could be made around the opening up of certain amounts of appropriate financial data. [What legislation regulates which bodies are subject to scrutiny from the Audit Commission or its appointees?]

With university based research benefiting both from the award of project based research funding largely from the UK Research Councils and histroically from block grant funding from HEFCE (http://www.bis.gov.uk/assets/biscore/science/docs/a/10-1356-allocation-of-science-and-research-funding-2011-2015.pdf)

Cultural and heritage data/metadata, eg in the context of the BBC Public Commons initiative, or the JISC UK Discovery project, should be made available in an open way when the production of the metadata is covered at public expense.

The complementary PDC consultation makes great claim regarding the definition of public task information insofar as public bodies are covered by PSI regulations relating to the reuse of public sector information (paras. 2.7-2.10). However, the phrase "public task" and "public task information" make no appearance Making Open Data Real consultation, which I find confusing? When trying to define policy that will determine which bodies must release (or make accessible) what public data, will there not be some interaction or crossover between that policy decision and the public task definition?

****2. Where a decision is being taken about whether to make a dataset open, what tests should be applied?
If data is part of a formally defined data burden, should that data burden be tiered in terms of openness requirements, for example along lines of:
- open on submission to the centre;
- open following embargo period and subject to checking by the centre, but with the presumption that it will be opened;
- not open;

Where data is FOIable, that may be taken as evidence in favour of presumed openness. If data is regularly requsted via FOI, it could be made available in open form as a matter of course in order to reduce FOI overheads in the future. When data is released via FOI, it could be made available via an open data site in partial fulfilment of handling the FOI request. When responding to FOI requests for data, the process required to obtain and release that data could be captured and compared with the actual processes relating to operational and administrative use of that data in order to identify whether an open data tap can be introduced into the current data process to open it as a matter of course, or release it efficiently in response to an FOI request.

As the major producer and consumer of public data, public bodies are well placed to benefit from more open public data. "Publicness" and "openness" both help make data accessible for use within and between public bodies, we well as reuse by third parties; accessibility is also improved by timely release of data, and the publication of data using open standards and formats.

Consequences of making data open should also be considered; for example, once released, will there be continued access to regular updates of the data using the same format. (If the data is released sporadically and with inconsistent formats, services that automate the regular collection of the data are not really viable).

To what extent should data requested in servicing questions raised in the House of Commons or House of Lords to Ministers, the answers to which will presumably find thier way into the public record? For example, obligations placed on particular institutions (eg prisons) for data relating to programmes running by those institutions that need to be supplied at short notice in response to Parliamentary questions raised in the House of Commons or the House of Lords.

Data collected by regulators (Ofwat, Ofcom, CAA, Ofgem, ORR etc) and used as part of public reports should be made public as a matter of course.

As far as operational data is concerned, good arguments can be made (eg on planning or mitigation grounds) for making data open relating to the status of (criticial) infrastructure: for example, data used by Railtrack, operators of critical infrastructure, status information eg from BT on broadband netwrok status, or grid status alerts, live travel data and scheduled timetables. At the current time, there is no easy way to check the status of crtitical infrastructure, eg in the case of an outage (eg as discussed in http://blog.ouseful.info/2011/10/19/to-do-critical-infrastructure-status-maps/).

****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?

Where work must be done that does not represent value for money (what would an example of this be? Having to get data into a form the public body would never use?), it may be appropriate to consider the amount of value that is added in processing the data that the requester might otherwise be expected to add, for example as just reward for the cost of processing that data. If the raw data is open, and the requester asks for processed data, it may be appropriate to give the raw data away freely but charge for the value add of processing it that the requester seeks to exploit in the course of their business? However, there will also be a tension between people who want to gain access to a small amount of data, either for personal use, research/innovation purposes, and companies who make use of that data in volume as part of a business. In the latter case, we might expect some payment for use of the data once the business is operating, although it could be argued that if the business is profitable, there is a return built in through taxation.

A balance may need to be struck based on the number of independent requests that are likely to be received for a particular data set and the use they wish to put it to. If N requests are made for the data, and all N parties need to do the same work cleaning or processing the data in the same way, that is obviously inefficient. It may be that third parties process and repackage data, for a fee. But the question arises - if data as published is not fit for use by third parties, is it fit for use by the first (producer) or second ('official consumer') party, or has the data been produced solely in response to some openness criteria, and not because the data is actually used for anything?

The ability to save cost elsewhere in government may also be an issue. For example, local authorities who make disbursements to care homes need to mitigate against fraud by regularly checking death reports, often through the purchase of commercial death registers or by checking the local newspaper's death notices. Whilst a cost may be associated with signatories of death certificates ensuring this data enters the public body data chain in an accessible and open way, it may well save costs in multiple other areas of government.

Where data is processed and released in exchange for a payment, would it also be possible for the raw underlying data to also be made available for free so that third parties can, at their own expense, carry out the required processing if they can do so for less overall cost than piecewise purchase of data from the public body?

****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?

If an organisation is subject to FOI requests, or data it produces and returns as part of an official data burden may be requested through FOI requests, it should be in scope?

Analysis of data processes associated with fulfilling data burden requirements might provide a basis for identifying where in a data process data might reasonably be made public and open.

*****5. What would be appropriate mechanisms to encourage or ensure publication of data by public service providers?
If data related FOI responses are published via open data sites, the open data site can become a repository of commonly requested data and help identify which processes might benefit from releasing open data as a matter of course.

Where public data is reported as a matter of course by the local press and in the local interest, (for example, court reports, planning notices, traffic notices), public bodies might be encouraged to publish the corresponding data in an open way in order to facilitate the local dissemination of that information. Note that much of this data is transitory/may only be relevant for a limited period. In this case, we need to consider: whether there is a public interest in making the data publicly available and open on an archival basis, or not providing archives per se, but responding to requests for archival copies of data; the extent to which third parties can archive/aggregate such data and continue to make it available; whether there are privacy reasons for not supporting archival access (for example, court reports in local newspapers have a "short memory").

Are there guidelines available that cover the interactions between things like:
- data eligible for release under FOI;
- data that may be redacted on grounds of Data Protection Act
- data covered by Database Right or data that is covered by copyright
- data released through National Statistics
(http://www.legislation.gov.uk/ukpga/2007/18/contents)
- reusable public sector information
(http://www.legislation.gov.uk/uksi/2005/1515/contents/made)

Analysis of data burden reporting process might identify appropriate points at which data can be made open as part of the process. For example, reported data may be posted to an open data site from where it is collected ("pull reporting"). See also: http://blog.ouseful.info/2011/03/18/open-data-processes-taps-query-pathsaudit-trails-and-round-tripping/

And as I responded to the PDC Engagement Exercise, [o]ne particular class of data that interests me is data that is:

1) reported by a local organisation to a central body;
2) using a standardised, templated reporting format,
3) and that is FOIable either from the local organisation, and/or from the central body.

For example, in Higher Education, this might include data on library usage as reported to SCONUL, or marketing information about courses submitted to UCAS.

It can often be hard to find out how to phrase an FOI request to obtain this data as submitted, unless you know the type of reporting form used to submit it.

What I would like to see is the Public Data Corporation acting in part as a Public Data Exchange Directory, showing how different classes of public organisation make standard (public data containing) reports to other public organisations, detailing the standard report formats, with names/identifiers for those forms if appropriate, and describing which sections of the report are FOIable. This could also link in to the list of local council data burdens, for

example (http://www.communities.gov.uk/… and/or the code of practice for local authority transparency (http://www.communities.gov.uk/…)

The next step would be to introduce a pubsub (publish-subscribe) model in the reporting chain for reporting documents* that are wholly FOIable. This could happen in several ways:

A) /open report publication/ – the publishing organisation could post their report to their opendata reporting store, and the consuming organisation (the one to which the report was being made) would subscribe to that store, collecting the data from there as it was published; third parties could also subscribe to the local publishing store and be alerted to reports as they are published. If co-publication to the central organisation and the public is not appropriate, the report could be witheld from public/press consumption for a specified period of days, or published to the press but not the public under embargo.

B) /open deposit/ – the publishing organisation publishes the report/data to an open deposit box owned by the central organisation which is receiving the report. After a specified period of time, the report is made public (ie published) via that central deposit box.

C) /data corp in the middle/ – a centralised architecture in which local organisations submit public reports to a Public Data Exchange, which then passes them on to the central body to which reports are made, and publishes them to the public, maybe after a fixed period of time.

The intention of all three approaches described above is to provide an open window onto the reporting chain. At the current time, open public data tends to be data that is published via a separate branch “to the public”. In contrast, the above approach suggests that public data publication acts as a view onto all or part of the data as it goes about it’s daily business being published from one organisation to another. That is, public data publication becomes a “tap” onto a dataflow/workflow process.

If one of the desires for data exploitation is to help introduce efficiencies as well as reuse in data related activities, third parties need to be able to work with data as it currently used.

The extent to which FOI can be used as a lever for obtaining data releases should not be underestimated (foer example, in accessing research data http://www.jisc.ac.uk/publications/programmerelated/2010/foiresearchdata.aspx). When framing right-to-data legislation, or modifying the current FOI legislation, care should be taken not to lessen the scope of what data may be currently requested by this route. With changes in funding models to the universities, and the possibility of HEIs entering the private sector, to what extent will research data continue to be subject to FOI requests, eg in the case of a private university operating publicly funded research programmes?

***How will we ensure that Open Data standards are embedded in new ICT contracts?
By providing a test suite as part of the contract that include tests such as running data import/export/query operations against centralised validation services.

***What is the best way to achieve compliance on high and common standards to allow usability and interoperability?
Require data reporting to proceed through open interfaces or interfaces where public data taps can be applied. Released data should be authentic, and representative of data used as part of a

public body's activities or reporting duties rather than data that is produced purely for release on an open data site.

***How would we ensure that public service providers in their day to day decisionmaking honour a commitment to Open Data, while respecting privacy and security considerations.
Take a lead from open source software projects and publish requests via an issue tracker, that can show when an 'issue' was raised, what it's current status is, and how it was resolved. Related approaches include services like WhatDoTheyKnow or GetTheData

***How should public services make use of data inventories? What is the optimal way to develop and operate this?
If we distinguish between datasets, queries on datasets, and reports/data view generated by queries on datsets on the one hand, and data burdens on the other, we can start to map out how queries are used on datasets to generate reports that fulfil data burden requirements. That has the benefit of making the data burden fulfilment process more transparent, as well as contextualising both the way those reports are generated (through exposing the queries) and the original data sets used as a basis for creating reports.

***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?
One rule of thumb is that the data should be "good enough". The question then arises, 'good enough for whom?'. If the data is released and never referred to, its quality is irrelevant as regards the non-existent on-users, although it may signal problems elsewhere. If data is used by a third party and found to contain errors or omissions, the question arises: does the publisher also suffer from those some lack of quality issues (and if so, how are they handling them?); or are they using a different data set as part of the process that the released dataset relates to (and if so, why isn't that data being released?)

There are different levels of cleanliness we may associate with data: a major issue in many datasets relates to the use of inconsistent labels to refer to the same entity (something that can be addressed by using universal persistent identifiers). Character set encodings can also cause problems, especially where it is hard to identify what character sets are used within a file.

***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?
As I understand the current situation, public body reports often produce summary tables and as part of transparency requirements, release as public data raw datasets that are used to generate those summary tables. In such cases, the query used to generate the summary table from the raw data should also be published. The transparency does not come from releasing summary tables and saying "it summarises that pile of data". It comes from saying - here is the summary, and here is how it was generated from that data, allowing the observer to check the assumptions of the query, redo the analysis, and so on.

Using services such as Google spreadsheets or Zoho spreadsheets, it is possible to provide a preview view over the data contained in a dataset made available as a simple CSV file (this approach is taken on some datastores). It is also possible to use services such as a Google spreadsheets as a database, and so provide a certain level of intermediate developer access to the raw data as if read access were made available to the database that sourced the released data (eg http://blog.ouseful.info/2010/11/19/government-spending-data-explorer/). A range

of powerful hosted statistical analysis and visualisation tools are now available that can also provide a user interface layer over over data published in such environments ("analysis at the point of delivery"). For example, the popular R environment can provide an online statistical analysis UI to online hosted datasets via services such as http://www.stat.ucla.edu/~jeroen/ggplot2.html or http://www.rstudio.org/docs/server/getting_started These tools provide an intermediary step that allow interested parties to explore datasets in situ. Recent developments with the Linked Data API (http://www.epimorphics.com/web/tools/linked-data-api.html) offer similar capabilities, including the ability to share persistent URLs to queries that are applied to public Linked Data stores such as those hosted under the data.gov.uk umbrella.

****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?
Allow free access to public data for personal, research, social enterprise and SME commercial research/development purposes. If the service using the data ever becomes popular, worry about how to charge for it then...

To what extent are services like the Digital Curation Centre capable of providing advice to public bodies faced with managing significant public data archives locally?

Is there a role for Technology Strategy Board (http://www.innovateuk.org/) initiatives such as Knowledge Transfer Partnerships, in which data wranglers work with public bodies to find how data is currently used in government and explore ways of complementing that with open data projects, or the Small Business Research Initiative (SBRI) http://www.innovateuk.org/deliveringinnovation/smallbusinessresearchinitiative/whatissbri.ashx in which data credits or similar benefits are provided to small companies looking to developer commercial or third sector products or services using public data.