Response to the UK Government's "Making Open Data Real: A Public Consultation"

EnAKTing Project, Web & Internet Science Group, University of Southampton

This document sets out the Response from the EnAKTing group of the University of Southampton to the UK Government's "Making Open Data Real: A Public Consultation" [1]

EnAKTing is an EPSRC-funded project dedicated to solving fundamental problems in achieving an effective Web of Linked Open Data. We have good hands-on experience of working with data from data.gov.uk, and turning it into linked data. We provide visualizations and applications supporting a range of benefits. For example, enabling the public to gain insight as to the effectiveness of public services, such as crime, education, health, in their locality. More background on our work and the types of application built can be seen at [2].

The Response

We will not provide responses to all questions but rather to those for which we have particular expertise.

With regard to the questions of **setting transparency standards** we have a number of specific and somewhat technical responses that we believe it is important to state. The adoption of new Open Data standards must be included in the guidelines for new ICT contracts in order to enforce their application. A national monitoring organization needs to provide guidelines for those writing ICT contracts to ensure that those guidelines are followed. This will serve two purposes: first, it will provide incentives to publish even more open data; second, it will ensure that funded ICT contracts provide adequate levels of data quality.

When it comes to interoperability, it is important to recognise that interoperability can be achieved at different levels: data representation, resource identity resolution, etc. Moreover usability can also be influenced by a number of factors for which, the adoption of common standards, technologies, and workflows can provide a solution. Open data initiatives are now taking place in a growing number of countries around the world. Data interoperability is a problem that will affect potentially all of them. Eurostat already publishes data about more than one country, there is obvious merit in adopting common standards and procedures – if these are sufficiently agile and lightweight.

The first step toward usability and interoperability is to provide authoritative references for core data. Location [3] is already recognised as essential when making sense of Public Sector Information (PSI). Ordnance Survey provides an example of the benefits achievable by providing authoritative stable URIs for geographical regions.

Within EnAKTing we have provided services that can support a distributed publishing mechanism of linked data re-using already available PSI geographic URIs from Ordnance Survey [4]. We think these sorts of services will become central in the future as we seek to produce an integrated data landscape.

The same approach to distributed publishing/discovery/repurposing workflow for PSI can be further developed by asking local and national governmental bodies to manage an authoritative and stable source of URIs that can be collected and aligned

as has already been done for some of the data published by *.data.gov.uk (e.g. statistical data, education, etc.).

URIs for geographical regions have already been minted by different parties (e.g. OS, data.gov.uk) but not all of them are being maintained. The questions to ask here include: who will physically maintain the resolution of these resources and services?

Time is another dimension of many PSI data sets. In this case the URI space is complex if we want to take into account temporal intervals for real case scenarios. Standard procedures to define temporal resources with namespaces (that gives a context to the annotated data) must become common practice.

When it comes to the question of **how to ensure collection and publication of the most useful data**, an infrastructure is needed to ensure proper monitoring of user demand for data – and that once elicited the demand is met. This will make it possible to discern the most useful data. Once this is done, this information needs to be made available to the key publishers and service providers so that the life-cycle process of publication and consumption can commence.

The key actors in the OGD ecosystem include:

a) The general public who want to consume and view data that is useful to them,
b) The publishers that have the authority to make available the data in open data format
c) The service providers that will build services using these data and who also have the capability to link the general public with the publishers.
d) Public and private corporations including local and national Government will all be consumers of the data.

Tools, and tool support, need to be built by members of these groups to facilitate the following:

a) Measuring the inherent 'value' of the data, (through user demand, user feedback, and how these can be used to rank and categorise data)
b) Tools to enhance the data so that service providers/consumers can easily discern its value and quality (meaning, completeness, consistency, source, quality, temporal validity (i.e. freshness), accuracy, and provenance).
c) Tools to raise awareness of the availability of the data.
d) Tools to monitor the life-cycle process of the publication and consumption of the data, the vocabularies used in the data, and the applications that make use of them.

In this marketplace, there are 4 main elements we need to consider:

a) The data itself,
b) The catalogues that hold the data and/or references to the data,
c) The applications built on top of the catalogues such as dashboards with catalogue statistics, and
d) Services and applications that make use of the data Infrastructure.

Such community-based workflow must be supported by well-defined policies that can handle robustly the evolution of the data asset. One of the challenges in using new data sets is that they need to be clearly understood before they can be integrated into the current ecosystem of data sources. This issue is even more evident for OGD,

where different departments release data independently, describing data with for example different geographical and temporal granularities. Thus, the creation of glossaries and indices that explain the terminology used in the data is essential.

Similarly, open data benefits from exploratory interfaces over it. Users can explore aspects of the open data that they are familiar with, and discover related open data quickly.

On the question of **how or whether the government should publish high quality data**, it should be recognised that data quality has multiple dimensions including:

a) Accuracy - are facts in the data set correct?
b) Intelligibility - can the data be understood by the general public?
c) Referential correspondence - are resources identified consistently and without duplication?
d) Completeness - do you have all the data you expect?
e) Modelling granularity - does the modelling capture enough information to be useful?
f) Attribution/Provenance - can you tell where the data came from?
g) History - can you tell who has edited the data and when? etc.

It is not likely that all data will meet the highest standards on all quality dimensions. The government will need to balance the provision of “high” quality data and timeliness of data provision. We would recommend that the government have a process in place that takes account of this balance in the act of data provisioning.

An act of data conversion and publication assumes a repurposing/enhancement of that published data. Public service data providers need to use services (such as e.g. http://sameas.org/) that can be used to connect, enrich and enhance data.

A central portal should not be unnecessarily *dirigiste*. Given that Web technologies are premised on distribution and decentralisation, it seems more appropriate that data be released close to its originating body, but catalogued centrally. This carries with it a risk that the data may outlive its originating body (in the event that departments merge, etc.). Therefore, we recommend that all data should be released with a contingency migration and curation plan.

On the question of **whether it is more important for government to publish a broader set of data, or existing data at a more detailed level**, we have found that much of the value in data lies in the ability to combine multiple data sets; simply publishing broader/more detailed data does not necessarily improve the utility of the data. It’s more important to publish data that is better linked to existing data sets, increasing the level of reuse internally.

The government can stimulate innovation by leading by example; look to imaginative reuse of open data within government and creating “incubation” funding and policies for entrepreneurs that want to exploit this data asset. Government should actively promote an application store model for the provision of its public services.

A Major Question Missing From "Making Data Real"

A major question missing from the consultation questions is "**How does Government Get Value from Open Government Data**", delivering on the promise of "Do more with less".

To realise the full value from the whole exercise of identifying, (re-)formatting and publishing its data, government has to plan to consume the data being provided, rather than regarding the whole process as ancillary to normal IT systems and internal processes.

Firstly, it should be a policy that government moves towards IT systems where the data being processed internally is actually the same data that it is publishing. This is a long-term objective, needing some changes to business processes and requiring appropriate specifications in the procurement of IT systems.

Secondly it should be policy that where government is accessing data from other departments and offices, the source of choice should be the Open Data version.

Without these moves, the Open Data activity will be severely weakened, as the only motivation will be the political agenda (Transparency), and the "pull-through" from a small but increasing band of data consumers in the commercial and non-commercial sectors. The whole activity appears simply a drain on the publisher's budget, without any related efficiencies. If this continues the danger is that the quality of the data is compromised. In this scenario there are few users to provide quality checks, and this of course leads in its turn to fewer users of the poorer data. We are beginning to see the effects of this at data.gov.uk, where many of the datasets have not been refreshed, as they should have been.

References:

[1] Making Data Real: http://data.gov.uk/sites/default/files/Open%20Data%20consultation%20August%202011.pdf

[2] EnAKTing: http://www.enakting.org also an example application from the research http://apps.seme4.com/see-uk/

[3] Location: http://location.defra.gov.uk/

[4] EnAKTing Geoservice: http://geoservice.psi.enakting.org