Introduction

This consultation response has been framed by a group of expert practitioners in the publishing, use and re-use of open government data. This response is the result of a one day workshop discussing the issues raised by the consultation. Our aim is to magnify the impact of our response by reaching a collective view, that this document records. We have limited ourselves to responding only to those parts of the consultation where our technical expertise and practical experience are most relevant. Therefore we have concentrated on the questions about standards, meaningful data, innovation and the government setting an example.

The UK government aims to be the most transparent in the world. That commitment is welcome and the government's determination evident. Our experience is that the use of Linked Data standards and technologies are particularly beneficial in achieving the aims of transparency. In answering the questions below we have set out to explain why, give examples of successes to date and highlight the opportunities we see in the future.

Ian Davis, Talis Group Ltd
Paul Davidson, LeGSB
Richard Goodwin, TSO
Martin Merry, epimorphics
Dave Reynolds, epimorphics
Bill Roberts, Swirrl IT Ltd
John Sheridan
Buddug Williams, Talis Group Ltd

27/10/2011

Setting transparency standards: what would standards that enforce this right to data among public service providers look like?

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

Transparency depends on meaningful data. That means data needs to be published with its context included. We need to know what the values in a dataset mean, unambiguously, in order to process and use that data. The benefits of transparency hinge in large part on the ability to compare information from different sources, quickly and reliably. The only way to do this is by using common standards. Without standards, there can be no meaningful data. Without meaningful data there can be no transparency.
There should be a presumption that we publish what data means, alongside the Open Data made available. This will give confidence to users of the data – a prerequisite for those looking to use that data in commercial applications. There is significant value and relatively little additional overhead of doing this in a formal way, using Linked Data standards.

To have meaning, data needs to be linked to something - a shared term, definition or concept. We would differentiate between linkable data, that is data which uses defined terms and reference data (such as Sedgemoor District Council's spending data), and Linked Data, where terms are defined explicitly, using standards, baked in as part of the data, with the aim of being shared or aiding comparison with other data. Linkable data is an important step on the road to Linked Data.

Linked Data is the best approach for publishing meaningful data that can be easily compared. This is why Linked Data standards are ideally suited to supporting the aims of transparency. Data is published with its meaning included, through shared concepts and relationships. This meaning can be shared with other people - they can use the same concepts and relationships in their data as you have done, or link their concepts (the things their data is about) to yours. Using shared reference data, about, say, organisations, types of spending, administrative areas or time periods, means that data can be compared more easily. The data can be easily queried across the web, through APIs (the Linked Data API, used for organograms and elsewhere, is particularly powerful and flexible) or through a query language called SPARQL.

We were pleased to see the 5 star model outlined in the consultation document. The 1 or 2 star approach is not good enough for achieving the aims of transparency. Just having the data under an open licence or in a machine readable format is not enough. It also is important to note that the benefits to publishers mount as you move up the star scheme, as well as to consumers. Only with 4 and 5 star data are the benefits envisaged from transparency likely to come to fruition, for both publishers and consumers.

Comparability is not always possible, even with some 4 star data. Linking data (the 5th star) is the key. For example, with the local government spending data, several local authorities used RDF and URIs – 4 star data. CIPFA expenditure codes, recently made available in RDF, provide a prism through which different local authorities spending data could be compared. We now need to encourage Local Authorities to reach for 5 star data using these authoritative reference data to link and bring comparability. Without comparable data, one of the main benefits of transparency to the local authorities themselves, benchmarking expenditure with others for similar items, or aiding smarter procurement, is lost. A comparatively small amount of effort, to link transparency data to a common standard, results in a disproportionately large benefit, from comparing that data with other datasets. To realise those benefits the government needs to actively promote linking data to core reference sets.

We recognise there is a cultural shift required for government to use Linked Data. Equally we have found the benefits of the Linked Data approach, in terms of responsible publishing of government data (with meaning, quality, provenance all included) dovetail with the concerns of many data holders. Data is less likely to be hugged, when it can be published really well.

The incentives to use standards, to link and share data, must be driven by benefits to the publisher of the data as well as to the consumer. There needs to be more emphasis strategically on the internal consumption by government itself of transparency data. Government is amongst those with most to gain from Linked Data.

The use of standards, in particular common reference data and common vocabulary, are necessary to achieve these benefits. There needs to be a mechanism for getting transparency data standards created and used. This requires a small but significant level of investment, co-ordination and leadership from somewhere in government. Experience says the sums of money involved are very small, there are lots of people willing to engage to support such activity inside and outside of government, but leadership is essential.

We are encouraged by the ICT Strategy. In particular the Open Standards Board and Open Standards Panel are welcome developments

We would suggest a number of carrots and sticks to encourage the use of standards.

The “carrots” - standards need to be:

- Genuinely open and free
- Visible and findable by data publishers and consumers
- Attractive to use with clear benefits to information holder. Incentives and benefits of standards are more important than mandating them, although both approaches may be needed.
- Easy to implement
- Developed and maintained collaboratively – those using the standard should shape how it is created and evolved
- Developed and presented in terms of business benefits / results. The standards need to deliver more for information holders than just helping an external developer to use Open Data – they can and should aid significant cost savings and efficiencies inside government, reducing duplication of effort.

The “sticks”:

- Public sector information holders should be evaluated on the basis of the savings / efficiency gains they have achieved through standards adoption and re-use. This needs to be measured and reported on as part of the organisation’s annual accounts, with a feedback loop in terms of the organisations budget and performance in subsequent years.
- When public bodies are told to publish data about something (eg spending, contracts or organograms), there should be a clear statement of what, exactly, should be published, with an expectation that everyone uses the same standards to do that, ideally (as in the case of government organograms), Linked Data standards. Unless the data has been published according to the standard, the requirement to publish that data should be deemed, “not met”. Without standards, everyone shares the pain of doing the work, but no-one gets the gain of comparable data.

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

Yes, there is a role for the government to establish consistent standards for collecting and publishing user experience information. The ambitions the government has for public services, set out in the Open Public Services White Paper, depend, to a significant extent, on data and large scale data aggregation, including user experience data. Open data is the enabler of choice.

We believe Linked Data standards afford unique benefits as they allow that information to be contextualised and combined. Alongside user experience data the government needs to collect and make available contextual data, in Linked Data form, to enable comparability of service performance and therefore service user choice.

The advantage of using Linked Data standards is that they work the way the web does – low cost distributed publication with a highly scalable infrastructure. Every publisher and consumer benefits from the network effects, as more data is added to the web of data. The use of URIs to name key concepts (services, service providers etc) means that people can make statements about those things on the web. High quality URIs for things, can enable the government to conduct web scale sentiment analysis about services or policies, using the whole web as an information resource about user experience, informing policy and practice as a result.

One powerful example of the power of high quality URIs is legislation.gov.uk. Since the launch of the service, people have made increasingly specific links to legislation on twitter, commenting about the laws that govern them and providing those resources with a context (someone's opinion). This trend is particularly noticeable on twitter – but the links to the legislation.gov.uk URIs could be exploited to develop a much richer understanding of society's and the economy's relationship to legislation.

We believe government should lead work in this area. It is a topic that greatly benefits from a co-ordinated approach, with planning and co-ordination from the centre. This is also an area where the government will most strongly see the benefits of large scale information aggregation, using the web, for itself. In general we strongly advocate that the government 'scratches its own itches' using Linked Data, focusing on internal as well as external use cases for large scale information aggregation and data processing, to enable Open Public Services and citizen choice.

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

For the purposes of this consultation response, we understand an information intermediary to be a person or organisation that takes (often raw) data from somewhere, refines it and provides the enriched / added value data, for others to re-use. Generally

there is a charge, applied by the intermediary, for the re-use of the higher grade data, which in turns helps to fund maintenance of the data they produce.

We believe an accreditation scheme to be unnecessary bureaucracy, particularly in an environment where ready, non-exclusive access to and re-use of data was assured. The power of web and open data is that anyone can be an information intermediary. If the intermediary links back to the public sources of their data, the community of data re-users can verify the accuracy and quality of the intermediaries work.

A key issue around intermediaries is that of trust (can I believe this data, can I use it in my product or service?) and the related question of quality. Not all information is equal – and open government data cannot be of uniform quality or accuracy, but the provenance of that data should universally be expressed, from the source or from an information intermediary. We would highlight the Linked Data approaches for provenance, quality and trust. That individual data points have a URI which de-references to some data, adds significantly to the assurance that can be given to consumers of Linked Data. There are Linked Data approaches for representing quality of data and even, at the high end, digital signing of Linked Data, that are significant benefits in this area.

Meaningful Open Data: how should we ensure collection and publication of the most useful data, through an approach enabling public service providers to understand the value of the data they hold and helps the public at large know what data is collected?

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

We think data inventories are a useful tool, for internal information management, for managing information risks, and for external transparency about what information a public body holds. Data inventories should concentrate on cataloguing datasets which are already public and those where it is conceivable that they could be made available if there was sufficient demand. There is value in putting information about unpublished data in an inventory, so we all know what we are missing, as well as what is currently available.

There are a variety of types of information and data inventory kept by government, from data.gov.uk and local government lists of data, to Information Asset Registers (IARs) and Publication Schemes. It makes sense, from the perspective of users, data holders and owners of data inventories, to look to consolidate the various initiatives around

information and data inventories – rationalising the various obligations. The INSPIRE obligations provide a catalyst for this too.

Data inventories should be developed using Linked Data standards. The technology is a good fit for managing metadata about datasets, and for facilitating distributed, interoperable data inventories. One key benefit is the facility to vary descriptions of dataset by type and to use common vocabularies, like DCAT, that enable interoperability between inventories. The codification of terms (such as the sensitivity level, personal data, contains 3rd party rights) into a concept scheme would aid the operation and interoperability of data inventories.

Many public bodies have large enough data holdings to warrant their own data inventory. This should be encouraged, as part of a distributed approach. Building a large central database of datasets is unlikely to be successful, either for data holders or for re-users. Data.gov.uk gas an important role to play as a central hub for a network of inventories. Data inventories should use Linked Data and should ideally be available as open source software. They should be distributed and interoperable. Datasets should be described using shared vocabulary, but not necessarily to the same level of detail.

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

It is very difficult for public sector information holders to know exactly what data is useful. Beyond clear cut cases (whenever transport data has been released in the world, new applications using that data have followed shortly afterwards) the government should follow both a pro-active release policy, of data it suspects may be useful, and respond to re-user demand.

There are four priority areas for proactive data release:

- Data that provides data for other data (often, by definition, this is reference data, but may also be common vocabularies)
- Reference data that reduces duplication of effort, so information is collected and managed once in government, and re-used many times.
- Data (performance, economic, statistical and scientific in particular) that supports policy decisions and aids the process of meaningful consultation about policy matters with business and the public. For example, if you were consulting on a Public Data Consultation, it would make sense to publish the data that supports the assertions in the consultation document.
- Other plausibly open data

The government should scratch its own itches. If one part of government re-uses data from elsewhere in government (either within the same department but from a different unit, or from another department), there is a strong likelihood that data will be useful to

people from outside of government too. Where possible, government should look to share data with itself, through publishing that data on the open web. This is both enabling for the wider economy and likely to be far more cost effective than point to point data sharing solutions inside government.

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

If government is doing something, or funding something (the provision of roads or schools say), then it should create and publish reference data around those things. This means creating URIs for those things, so government can publish related data about them (such as user experience data), and other people can relate their data to those things. The URI Set developed by Companies House for companies is a good example of this. Other reference data holders in government should do the same.

The publishing and maintenance of core reference data needs to be routine and assured. Reference data needs to be up to date. If it is based on snapshots, then it needs to be refreshed at predictable and regular intervals. Reference data that isn't maintained is of little use. Commitments to maintain reference data, such as obligations written into contracts with suppliers, are a powerful reason for others to trust, and therefore use, that data. This is essential for those who are looking to invest in product development and commercially exploit government data. Where government has backed its reference data with assurances over its persistence and quality (eg data from Ordnance Survey, or legislation from The National Archives), then others have used it with confidence for commercial applications.

Linked Data provides the ideal way of creating, maintaining and publishing government backed reference data and enabling its re-use. Taken together a coherent set of such core reference datasets would form a national information infrastructure which we believe would be a significant competitive advantage for the UK.

What data is collected 'unnecessarily'? How should these datasets be identified? Should collection be stopped?

Government departments should stop collecting data that is held authoritatively somewhere else (often somewhere else in government). Duplication happens all the time in government because of the lack of availability and/or trust in reference datasets.

For example, how many Statutory Instrument databases are there in government? Most departments have one, The National Archives has one, its contractor has one, the two Journal Offices at Parliament have one. Information is needlessly re-keyed and re-keyed again and again, with organisations needlessly spending money on creating and

maintaining their own systems. Another example is the SNOMED vocabulary, that was created for medical / health purposes. This includes a list of welfare benefits, that the SNOMED team find hard to keep up to date, but which they have included because there wasn't a definitive, up to date, reliable alternative source for the data. DWP could, and should, be such a reliable definitive source of a list to which SNOMED can simply link.

Duplication of data is deeply inefficient. It has happened in the past because the costs of co-ordination, to agree and use shared reference data, outweighed the benefits. Those costs are massively reduced by the web, and by the use of Linked Data in particular, with its emphasis on URIs for things. One of the advantages and benefits of Linked Data for government is that it systematically roots out duplication. There are likely to be significant cost savings to government from such an approach.

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?

Not all government data is equal and not all data can or should be of the same quality. In a world of varying quality of data, it is important to express the caveats associated with datasets, so re-users can make appropriate choices about how they use the data. Linked Data aids the publishing of datasets of varying quality attributes and with varying provenance. This kind of contextual information can and should be baked into the data, using Linked Data standards.

There is a good list of attributes of data quality here: http://answers.semanticweb.com/questions/1072/quality-indicators-for-linked-data-datasets

In terms of data polishing, having data which is sufficient for the purpose, is the goal. Re-users understand different datasets will have different quality attributes. The quality of reference data is particularly important. It needs to be regularly updated in order to gain trust to be used. For reference data completeness, timeliness of updating and persistence are key attributes.

We think there is benefit in expressing the quality characteristics associated with a dataset, as part of the data. This is another area where Linked Data standards can help. We have been in a world where the rate of inflation from ONS, or legislation from government, has been trusted. Now we are opening more datasets, many of which have lower quality attributes, it is more likely the data be used for purposes for which it is not fit. This is why we need to highlight lower quality attributes associated with data to help avoid accidental mis-use. One way of doing this is to express sources and methodology information (where the data came from). This is an area requiring further work in terms of standardisation, and leadership.

Government sets the example: in what ways could we make the internal workings of government and the public sector as open as possible?

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 is no reason why every table of figures (or even every multimedia data visualisation), in every white paper or green paper cannot be linked to the underlying statistical data the tables were drawn from. These statistics in turn can contain information about how the data was gathered and what statistical techniques were used. Using Linked Data we can relate policy proposals to the data that supports them, linking between textual information and structured data.

These approaches work best when the publishing of data is close to the people who create and manage the data. Excessive centralisation of publishing of data will be a bottleneck, blocking or stifling innovation.

Where the data is published and where the services are that people use to access that data (portals), are different considerations. We can envisage distributed publishing of data with a mixed economy of a flagship central portal (such as data.gov.uk) and more specialised user centred portals in sector areas, or by department, or commercially operated portals. The relationship between dataset and portal should be many-to-many - one dataset can be accessed through many different portals, each portal can provide access to many different datasets on the web. It is important that the portal is not the only way to access the data, as we have sometimes seen from government in the past. Open access to machine readable data enables the many-to-many approach.

We already see a mixed economy of access points to data. On the web, data can be published locally but indexed and used by people everywhere. In such a distributed environment, there is significant value in the metadata layer, how the datasets are described. There is also value in portals which help to document data, associating the machine readable version of the data to the information displayed to the user on screen. Examples of this close association between user interface and underlying data including the organograms and legislation.gov.uk. We know there is a wide variation of requirements of public sector information holders, from heavy duty / industrial data publishers (such as OS, ONS), with considerable expertise, to local authorities just starting to publish core transparency data.

There are lots of options available to the public sector for hosting government data as Linked Data. Organisations like the Environment Agency, Ordnance Survey and even some local authorities such as Litchfield, have shown this can be done at relatively little cost. There are fully managed hosting solutions supported by professional publishing services, through to options for hosting your own Linked Data, using your own ICT. Some providers have offered hosting of Linked Data free of charge to local authorities, for example.

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

We would suggest the following factors, for prioritising datasets for publication:

- The extent to which the dataset aids transparency or participation in democracy. This is particularly important with local data.
- The extent to which the dataset might fuel innovation, based on comparisons with other jurisdictions (e.g. transport timetable data begets new and innovative applications wherever it is released in the world)
- The extent to which the dataset adds to the network of Linked Data. Some data has far greater network value, because it can be linked to from many other places, and therefore disproportionately benefits the network as a whole. The web is as much about network value as it is about the value of individual resources (pages, services, datasets).
- The extent to which the dataset has been requested by third parties. If someone outside of government asks for a dataset, the presumption should be to release it, as the data is likely to be beneficial to others too.
- Where government is either creating a new tool or communicating in some way, the data underlying that tool or communication should also be available. The act of building the tool or the service, indicates that there is an identifiable need for that information, as you are meeting a demand. The presumption should be that there will be a corresponding data need. Applications should be built on top of open APIs. Linked Data can make a significant contribution.

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

There is no either / or choice between breadth and depth of data to release.

The government should prioritise:

- Publishing existing open data using Linked Data standards, to enable the network effects. This would enable the data already made public to be more easily found and better exploited
- Publishing reference datasets, to reduce duplication, with guarantees about persistence. It is very hard for others outside of government to mint sustainable, trusted URIs for those things the government controls (pedigree is hard to gain).
- Prioritise data which can be re-assembled for another purpose (eg boundary change data allows other datasets to be remixed, which in turn allows us to understand world through a different lens).
- Make the data that is published as close as possible to what is held - rather than producing lots of cuts of an underlying dataset, when that underlying dataset could be made available.

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, we think there is a role for the government to stimulate innovation in the use of Open Data, both from the perspective of a creator and publisher of that data and as a consumer.

The government should fund demand, investing in solving its own challenges, in terms of policy making, or public service delivery, as a data consumer, using open government data. This necessarily involves innovation – government has many significant data use challenges and much to gain from linking its data - but it is not about picking winners. Rather government should identify its needs (eg “we would like to combine information from these two or three datasets”, or, “we would like to mine this data”) and go to the market with a requirement, using Open Data as a platform. There is significant potential for government as a large scale user and re-user of its own Open Data, adding to the pull for government data, commissioning and buying solutions that use the open data made available by other parts of government, as well as providing the push for more government data to be released.

The government should ensure the enablers for innovation, such as simple and clear licensing conditions, are in place. Issues such as non-exclusive, open licences, are enormously important – licensing policy is one of the key enablers of Open Data and the Open Government Licence by default. Where commercial licensing is in place for government data the government should press towards simplification and openness.

The government should facilitate communities around its Open Data. Many of the most valuable public datasets are large and complex – the insight which has gone into the collection or creation of that data and its ongoing management, is very beneficial in terms of its re-use. This owner/producer’s understanding of their data is potentially very valuable to others. For example, the legislation database available from legislation.gov.uk is a complicated and sophisticated dataset. The National Archives has worked directly with several re-users, to help them best understand how to make good use of the data available, developing a community of re-users around the data. Initiatives like Linked Gov, funded by the Technology Strategy Board, are important, as they recognise that communities can help each other with government data, whilst providing routes for data owners also to be involved. The government can help drive these types of example forward by initiating or participating in “open innovation” projects with open government data, as producer and consumer.

The government should ensure there is high quality reference data, with a commitment to maintain those reference datasets as Open Linked Data, on an ongoing basis. Many of the more interesting innovations with government data come when that data is linked and combined – something that Linked Data technology makes far easier to do. To enable this linking, there needs to be a spine of core reference data as an enabler for innovation by others. For example, if somebody wanted to build a website that allows other people to comment about schools, say, or activities for children in a locality, they would need reference data about the schools and a way of referring to the local area. There are obvious reference datasets, around key named entities in the public sector,

such as schools and hospitals, as well as things like time intervals and locations (addresses, roads, postcodes etc).

There are several examples we know of where the absence of a commitment to maintain reference data has curtailed commercial exploitation. One example is the Linked Data version of Edubase at education.data.gov.uk, which unlocks the value of that dataset. We know this would have been commercially exploited by now if there had been a commitment to keep that data up to date. The development and maintenance of a spine of high quality reference datasets, that can be used in Linked Data form, as an enabler for innovation, should be a key task for data.gov.uk and other reference data providers in government (Ordnance Survey, Companies House etc). Data users need to see that there is a sustainability strategy in place for the data they link to – people need confidence (that the data is going to be around, updated and refreshed) in order to invest, either money or time, in government data.