talis® consulting™
World leading expertise in
Linked Data and the Semantic Web

KASABI™

talis® aspire™
Teaching and Learning, Connected

HM Government

Talis Group Response to Making Open Data Real: A Public Consultation

Thursday, 27 October 2011

Talis Group Ltd
43 Temple Row
Birmingham
B2 5LS
United Kingdom

+44 121 237 6034
http://talis.com

About the Consultation

In August of 2011 the Cabinet Office and Department for Business, Innovation and Skills issued two public consultation papers, one entitled Making Open Data Real: A Public Consultation and the second entitled A Consultation on Data Policy for a Public Data Corporation.

These consultations have been widely discussed by those interested in Open Data for transparency, efficiency and innovation.

Talis Group is one such interested party and this document forms the response from Talis Group to the Making Open Data Real consultation.

The consultation focusses on a right to data and what structures and mechanisms might need to be in place to make such a right effective. The consultation covers the areas of

- Key terms
- An Enhanced Right to Data
- Setting Open Data standards
- Corporate and personal responsibility
- Meaningful Open Data
- Government sets the example
- Innovation with Open Data

The deadline for responses is today, 27 October 2011. We believe that as this deadline passes and the responses are analysed by both Cabinet Office and BIS further discussion will be extremely valuable in helping form conclusions that are able to support real and meaningful open data for the UK.

About Talis Group

Talis Group Ltd is company with a more than 40 year heritage in data and information science. This response has been brought together by the group as a whole, with contributions from all three of our current business areas:

- Talis Systems Ltd
- Talis Education
- Kasabi

Talis Systems Ltd is a software and consulting business that works with companies and public bodies to help them understand and publish their data openly. We have provided substantial services and expertise to the data.gov.uk project and have provided training to many data teams within government.

Talis Systems currently hosts many of the data.gov.uk linked data services.

Talis Education is a young business focussed on software to support learning in higher education. Its initial product, Talis Aspire, is used by many leading universities in the UK and worldwide.

Kasabi is a new online data business. Kasabi's goal is to make data as easy to use, and as easy to publish as possible. Kasabi wants to help people unlock the value in their data, whether that means making it freely available — in order to drive innovation — or to explore more commercial models.

Talis Group has been involved in many aspects of the Linked Data and Open Data movements. We have provided support for various developing data communities and have been asked to speak with many organisation and at many events. Talis Group were involved in bringing the recent Semantic Technology and Business conference to London.

Talis Group funded the legal work required to develop a data-specific Open Data license. This is now maintained by the Open Knowledge Foundation as the Open Data Commons license.

The Talis Group businesses all make extensive use of open data and understand many aspects of the business, legal, social and technical requirements of making use of data.

Summary of Response

The consultation questions provide a good framework for structuring the conversation around how best to make Open Data real for the UK. We have provided specific answers to the questions in the sections following and felt a summary of the recurring themes would be useful.

We believe there is a great deal of opportunity presented by HM Government publishing data for re-use by individuals and companies alike. These opportunities fall into several key categories:

- Transparency
- Informed Choice
- Efficiencies
- Innovation

All of these agenda for open data are important and all have similar requirements in order to make them successful.

1 — Publish Data

Data that is published openly is far more usable than data that has to be requested. Often people won't know what to request or what might be available and often the time delay between requesting and receiving data is off-putting.

2 — License Openly

An ecosystem based on data requires certainty of licensing in order to make use of the data without fear. Provide clear and unambiguous licensing of all published data to support experimentation. This licensing must allow commercial exploitation of the data if we are to see investment made in new businesses.

3 — Remove Barriers

Use of data is often experimental; it is often an exploration to find an answer. That journey can happen much faster if there are fewer hurdles in the way. Any process that prevents direct and immediate access to the raw data should be avoided.

These criteria are common to all of the agenda that people pursue around Open Data and can be summarised as:

Give people unfettered access to the raw data to with as they please.

Responses to Specific Consultation Questions

Glossary of key terms

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

We believe the definitions for Dataset, Information and Open Data are fundamentally sound. We would not typically include unstructured data in the term, but this is very much a subjective point and we are comfortable with its inclusion here.

We believe the definition of Public Services to be too restrictive and believe it should not be limited to those established by statute, but rather to any service commissioned by a Public Body. There have been a number of cases of FOI requests rejected because the data is collected and stored by a private company on behalf of a public body and it is important that data in these cases can also be made available openly.

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

If a dataset is not explicitly protected by legislation (such as the data protection act for personal information) there should be a presumption that it will be made available under an open license. Innovation requires a low-barrier to access in order to facilitate experimentation and the more data we, as a country, can openly publish the more innovation we expect to result.

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?*

It is perhaps not unreasonable to expect that the requester of some data might have to pay some costs towards making it available, e.g. costs to collate the data from internal systems. However the costs to publish raw data in CSV and similar formats are very low, so a requirement to pay ought to be a very rare exception to the norm. Where cost is used to as a reason not to publish data this can be safe-guarded by requiring that an explanation and breakdown of the costs must be openly published in place of the data. This would provide for transparency in the process.

The presumption of free access to data is crucial to lowering the barrier to experimentation. Any process that requires registration, a statement of purpose or payment will restrict access to the data to those who are willing to tackle the process — however simple.

Ideally costs should only be incurred once, as once the data is published it should be available for anyone to use under an open license. This casts the model more towards "sponsoring" of the opening of datasets, in order to help them be published, rather than a system of ongoing payment.

There is also a troubling presumption in the question that we wish to raise. The owners/curators of data are often not best placed to assess value for money or potential of data. One of the goals of publishing open data is to encourage innovative and experimental uses. In this context, calculating a value of releasing any particular dataset is not possible.

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?

All providers of public services, whether public bodies or other service providers, should be required to make data available openly. The question should perhaps not be about which organisations, but more about the priority. Those organisations that interact with largest sections of the populace (education, transport, health, etc) should be prioritised.

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

For public bodies, there should be a statutory requirements. For other service providers, e.g. government contractors, this should be enforced by specific contractual agreements. If necessary there should be statutory requirements governing the right to data terms that must be present in public contracts.

Legislation should be used to ensure that only wholly transparent and direct costs of releasing data can be passed on.

An Enhanced Right to Data

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

All of the measures outlined in the document seem viable to help drive presumption in favour of publication. We believe it is also necessary to legislate, requiring data to be openly licensed by default, in order to prevent dilution of the presumption.

Additional measures such as assessing costs of FOI requests and a right to appeal are important, and should be built upon a basic legislative requirement for data to be open.

Access could be further improved by ensuring greater consistency of the use of the OGL. Some government data is available under OGL yet many government sites have not been updated to reflect this. This results in uncertainty over the terms of use and this prevents re-use of the data. Ensuring that the OGL is properly applied everywhere, will help clarify usage.

Further investment in data.gov.uk, or a similar catalogue, to ensure that all government datasets are recorded in a single directory would also promote greater access and usage. Making the catalogue data itself available openly is also important to support innovation in the discovery of data.

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?

We believe the best way to ensure a right to access and a right to data is through primary legislation that individuals or companies would be able to use through normal legal channels.

An independent body may be useful in this context as a promoter of the right to access and potentially an arbiter and ombudsman. We believe this should be in support of primary legislation rather than instead of.

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

We believe that the reviews currently in place provide adequate protection.

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?

A basic premise of making data openly available online without barriers to access presents little overhead. This is inline with 2 star data in Tim Berners-Lee's 5 star rating scheme1 for online data.

By licensing the data using an open license a market can develop in which the data with the most value can be converted, republished and enhanced. This is one aspect of innovation that opening the data can support.

Where possible public bodies should look to increase their "star rating", particularly for key reference datasets and this is particularly true for datasets that consist of important identifiers:

- geographic places
- lists of public bodies such as education and health organisations
- lists of public assets

[1]http://www.w3.org/DesignIssues/LinkedData.html

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

Standard clauses should be included in all government contracts, requiring open data publishing. Clear guidance on minimum compliance levels should be provided to all public bodies. Legislation could be used to ensure that rights to data are enshrined in supply to public bodies even if not explicitly within the contract.

Safeguards should also be put in place to ensure that exclusive access to public data cannot be granted to private companies in barter for services without sufficient transparency, oversight and redress.

Setting Open Data standards

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

Useful approaches to common standards on usability and interoperability are provided by both the Transparency Board Open Data Principles, and the 5-Star scheme for data publishing.

The approach recommended for public bodies should be clarified by providing best practice and code of conduct documentation. We would recommend that this follows the five star approach of publishing quickly and working to improve the quality of data and publishing approach over time.

The path towards 5-star data can be improved by continuing some of the early work pioneered by data.gov.uk in demonstrating how to publish high quality Linked Data. Continuing this effort to add more exemplars for key resources will help illustrate how the process can be followed in practice.

Data.gov.uk may be a useful central point for finding government datasets and commitments should be made to continue investing in the project. Ensuring that public bodies regularly update the directory to link to their latest datasets will be necessary to ensure the usefulness of any catalogue.

An accreditation scheme for datasets, e.g. to indicate their standing on the "5-star" scheme could similarly be held on data.gov.uk in order to motivate improvement in the publishing approach.

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

Commercial organisations use advanced techniques to measure usage of, and satisfaction with, their online services. Often these are absent from online services provided by public bodies. Government could provide best practice guidance for the data that should be gathered in order to asses the impact of online services.

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

We believe such as scheme would be an unnecessary overhead. There is a market advantage to proving the provenance and authenticity of data you use. If public data is published in a way that can be easily referenced and verified then such a scheme adds no additional value to the market.

Corporate and personal responsibility

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

We believe that making a top-down commitment in legislation will be necessary to ensure that open data is taken up universally.

Registering data in a central catalogue such as data.gov.uk also promotes the release of data by allowing the public to compare and contrast different public bodies. This is the transparency agenda applied to the publication process itself.

A rating system such as the 5-star measure provides strong motivation to improve due to the clear measure it provides and the perception of scoring lowly on such a scale.

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

Ensuring that an individual takes board-level accountability for the right to open data is likely to provide motivation and organisational focus within public bodies to ensure the agenda is met. A legal framework to protect privacy is already in place so only open data needs to be addressed in new legislation.

Whether that is the same individual responsible for data protection should be left to individual organisations to decide as they know best how to assign responsibilities within the organisation.

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

Sanctions are a necessary mechanism for ensuring that a right to data is upheld. There will be cases where the desire to withhold data is great and there must be sufficient sanctions available to ensure that data is released.

We believe primary legislation, providing enforceability by the courts, is the best mechanism for this.

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

A dedicated Sector Transparency Board could benefit all who interact significantly with the public sector, theoretically everyone yet in practice a smaller group with varied needs.

A Sector Transparency Board as the enforcement mechanism for a right to data risks not being able to scale to the challenge and not having sufficient sanctions available to ensure the right becomes embedded in public service culture.

Meaningful Open Data

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

The data.gov.uk project has already pioneered the way in providing a useful inventory for government datasets. The optimal approach is to continue and expand investment in data.gov.uk to create an efficient primary resource for finding government data.

The scope of data.gov.uk could be expanded in several ways:

- The range of organisations able to submit entries to the catalogue could be widened.
- The service could begin to relate government services, products and data sources
- The service could start to reference commercial enhancements to government data.

The availability of a central catalogue that adequately lists and promotes government data, and the services provided from it, can help build a market for data-based services and innovation.

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

Data publication should be prioritised based on a blend of

- what is being asked for by those outside of public bodies
- what is perceived important by data owners within public bodies
- what is readily available to release today

Blending these to produce a prioritised sequence of datasets should be the responsibility of each organisation, influenced by outside guidance where necessary. The prioritisation should be publicly available and published openly.

The overall inventory, or catalogue, needs to be comprehensive in order to be useful. To achieve that the objective should be that all public data is recorded in a single publicly available inventory.

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

We believe data collected and published by government falls predominantly into two categories; data on the performance of services and key infrastructure datasets.

Data on the performance of services should be routinely collected and published in order to provide transparency and accountability in the delivery of services as well as to support informed choice for the public choosing access to public services.

Key infrastructure datasets such as maps, road network, traffic and weather data as well as listings of public bodies and service providers must be maintained as they provide the key to connecting and making sense of other datasets.

Government regularly publishes a range of statistics measuring the performance of various public services. These are obvious candidates for priority publication as they reflect useful, well-maintained, easy to publish sources of data.

These datasets often refer to other entities such as schools, roads, hospitals, companies, services, etc. It is important to ensure that data about these entities is also made available. This core reference data provides a backbone that is necessary to understand and link together data from different sources, not just from government but also across the private sector.

Data that provides context to education could stimulate real innovation in the provision of education. The National Learning Directory and data collected by UCAS are not currently open yet could support significant innovation in education if made available with low-barriers to use.

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

Without a comprehensive register it is difficult to assess what data might be being collected unnecessarily. And, without public access to data, which may support unexpected reuses, it is difficult to assess whether the data itself is being collected unnecessarily.

Emphasis should be on ensuring access to as much data as possible in whatever form it can be released. Usage statistics can then be used to determine the priority of further spending.

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?

Aside from considerations around privacy, the emphasis should be on simply publishing data rather than polishing it. Quality can be improved over time either at source, or via third-parties that can compete to add value to government data. Without seeing what additional uses can be made of some dataset, it is harder to assess its "quality". That said, some basic quality indicators are important: i.e. the data should be regularly updated, and clearly described both in general terms and in structure

Highly quality data is an ideal, in the first instance its more important that data is published and licensed openly. Interested parties may then polish that data, enhance and re-publish it.

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?

We do not consider where data is held to be a key issue. The important considerations for publication of data are:

- How can the data be discovered?
- How can the data be accessed?
- What is the commitment to permanence?
- How will the data (and its discoverability) survive any organisational re-structuring?
- How will the data be archived to ensure its long-term availability?

It is possible to address these questions with either a centralised or a de-centralised publishing model.

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

At a national level, priority should be given to publishing basic data about key entities like companies, schools, hospitals, roads, etc. This includes basic informations about each entity such as its name and location.

These datasets provide a backbone against which additional local and sector based information can be inter-linked. This facilitates integration of data from different

sources and can help ensure that data is not isolated. This can reduce integration costs across government and the private sector.

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

As noted in the above answer, reference data on key public entities like schools and hospitals should be prioritised and published to ensure that there is a way to usefully link and collate data across different sources.

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?

Government has a rich treasure trove of data that spans many years. It covers key infrastructure such as lists of schools and hospitals as well as detailed information about many aspects of public services.

To understand the role government might play in stimulating innovation we must first accept that innovation is an entirely unpredictable venture that relies on experimentation and trial.

With that basic premise we can see that there are several things government must do if it is to stimulate experimentation in support of innovation.

The first thing required is simplicity. It must be simple to discover and access the large amounts of data available. This will require investment into services like data.gov.uk and also into the data teams throughout government so that data can be published efficiently in ways that allow it to be found and used.

We believe the second thing that is required is unfettered access. Not for privacy reasons, but simply because experimentation may take place in short bursts of activity as people have time. Any delay in accessing data due to registration or approval processes presents a barrier to immediate experimentation.

Thirdly, certainty of licensing. It is crucial that innovators and investors understand the intellectual property position of any data on which a business might be built. If licensing is unclear it is difficult to justify investment in the data as their is a clear risk that any business may be restricted in what it can go on to achieve.

Our education business, for example, is innovating within education and has clear ideas for services they would like to trial. To invest these would require unambiguous open access to:

- UCAS Institution and Course Data and UCAS codes
- The Skills Funding Agency's National Learning Directory

Opening datasets like these may require existing agreements and contracts to be changed as in many cases there have been exclusive agreements put in place.

— ends —