Making Open Data Real: Consultation Response

Response for and on behalf of Epimorphics Ltd

The overall direction and approach proposed in the consultation is very welcome and well supported by the evidence given in the annexe.

The opportunities identified for productivity gains, improvements in quality and outcomes, effective user choice and economic growth all require that it is possible to interpret and especially to compare the data sets released. Without any explanation or standardization of the data and the reference terms and vocabulary used (to identify services, locations, outcomes etc) then the data would be useless for such purposes. Without some coordination and standardization we can't compare and benchmark; without comparisons we cannot improve; without improvement transparency lacks point.

This need for meaningful, comparable data is *why* it is so important to invest in continuous improvement aiming for the four and five Star levels as described in 8.9. The Linked Data approach, which is implicit in that rating scheme, provides a means to create and document common terms and to attach the interpretation of those terms to the data, putting the data in context, making it meaningful.

From this perspective we offer the following overall recommendations:

- Commit to continuous improvement for public data data publishers in achieving the highest ratings in the Five Star scheme for Open Data.
- Prioritize the **open** publication of (a spine of) reference data that enables data sets to be compared and linked.[1]
- Actively invest in standards for data publication including common vocabularies, reference identifiers and quality information. The inclusion of an Open Standards Board and Panel in the ICT Strategy is very welcome but active development, not just passive selection, is required.
- Development, publication and curation of reference data sets and vocabularies will require some investment. While the cost of using linked data techniques is very low, compared to more monolithic IT standards, some investment is required to provide the foundation and support to make meaningful, comparable data publication the norm. A team should be created and *funded* to make this happen - to develop reusable vocabularies, to support publication of reference data sets, and to support their effective use. The investment required is very small in comparison to the benefits but to make *meaningful* open data the norm does require leadership.

A primary user of Government data is Government. At present Government departments needlessly duplicate and maintain their own copies and variants of many key data sets (legislation, spatial data) for several reasons - including uncertainty over the stability of data supply, license restrictions and cost. This leads to waste and lack of effectiveness in Government and Local Government. A key to ensuring that Open Data becomes standard practice and to create a sustainable eco-system of intermediaries who combine, refine,

[1]This affects the PDC consultation since several key reference identifier sets (addresses, spatial objects, admin/geographic regions) are entangled with the PDC.

analyse and present data is for Government to participate directly in the eco-system. From this perspective we offer the following two additional recommendations:

- Establish it as best practice that non-sensitive data flows within Government should be via Open Data standards. Data publishers should publish with the quality, timeliness and commitment to longevity that is needed to enable other Government bodies confidently to use the data rather than replicate it.
- Government should participate directly in the Open Data ecosystem by purchasing value added data feeds, data presentations and data-driven decision support tools back from the open market. Rather than develop expensive bespoke data processing over closed data, use market forces to drive down the costs by timely release of the "raw" data and purchasing the value-added services back.

An Enhanced Right to Data

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

Introduce a requirement that public bodies and providers of public services proactively publish data about the services they deliver.

Provide active, funded support for meaningful publication to ensure the data is used.

Make use of open data within Government. The presumption should be that inter-department use of data should be via open data flows. Currently there is substantial waste and inefficiency incurred by Government bodies replicating and curating their own versions of data (for a variety of reasons, but including license restrictions and lack of trust in longevity of data). The aim should be that open data is accurate, timely and maintained well-enough that Government departments and other bodies can use it for all non-sensitive data flows. Support for this presumption should be built into future IT procurements (see below).

In some ways the strongest way to create the presumption is to show it delivers benefits. Government should actively participate (see later) in ensuring that sustainable value added services are created over the data. It should track and monitor the usage of data and document the case studies which show the quantifiable benefits that result.

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?

A good step at least.

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

Yes.

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?

Publication of data is fundamentally a cheap process if the data already exists in usable form.

The aim should be to shift the use of data within Government to use open and lightweight

cross-Government standards internally and to implement non-sensitive intra-Government data flows via Open Data standards. In this way the additional cost of publication of non-sensitive data should be trivial.

There will be some costs involved in documenting and codifying the meaning and context of the data so that it is usable by people other than the data holder. This cost can be moderated by making documentation of data context and structure using open standards simply a matter of best practice anyway. A support team to help the transition to this style of best practice will be required and should be centrally funded.

There will be cases where data is held in non-releasable form, typically because it includes (directly or indirectly) sensitive information subject to data protection or national security constraints. The costs of anonymizing, aggregating or sanitizing of such data may be non-trivial. In those cases then case by case judgements of the balance of benefit v. cost must be made. We see no reason why existing processes such as via the ICO cannot make such judgements.

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

Support for Open Data standards, both for export and for import/federation of data sources, should an evaluation criteria for future data-related ICT contracts.

This applies not just just to information management software systems but to system architecture. When architecting systems to support information consumption (including public-facing web pages and decision support applications) there should be a clear separation between the data layer and the presentation/application layer, to provide for data reuse.

Setting open data standards

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

As argued in our introduction, the use of common standards is critical to making the published data meaningful and comparable, which in turn is required to in order to reap the identified benefits of productivity and service improvement. These standards include reference data and vocabularies, not just data formats.

First these standards need to exist. There needs to be a mechanism for getting these standards created and used. That needs some level of investment, co-ordination and leadership from somewhere in Government. Experience says the sums of money involved are very small and there are lots of people willing to engage to support this activity, but leadership is essential.

Then the barrier to use of the standards needs to be as low as possible. They must be tractable, fit for purpose and adaptable over time - large monolithic centralized standards would slow down publication. The web-based approaches to data representation and interchange, Linked Data, are a good match to this requirement. The standards need to be easy to discover and use - backed up by good web resources, tools and a community-of-practice support network. To be attractive and relevant to stakeholders they need to be collaboratively developed with effective involvement of the stakeholders themselves. The leadership and co-ordination role should be just that - leadership - not a self-contained standards development group.

Next the stakeholders must themselves see the benefits in publishing against common standards by seeing that the data is used and that new insights and improvements are possible as a result. This applies to not just external developers but to other Government users. A primary user of Government data is Government and re-use of Open Data should be rewarded.

Finally, there needs to be some feedback mechanisms to encourage open standards-based publication. The savings and efficiency gain through data re-used should be measured and used in evaluations. When bodies are instructed to publish data in a given category then that instruction should be accompanied by a clear statement of what is to be published and how, with an expectation that everyone should use the standard. Without this everyone shares the pain of publishing but without the standardization to allow meaningful analysis and comparisons then no one sees the gain.

Summary: coordination, lower barriers, maximize benefits, evaluate against publication and use.

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

Yes. The Government's ambitions for Open Public services will require data that must be aggregated from many places, and consistent standards for that will be needed for comparability.

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

The consultation does not define what is meant by the term *information intermediary* in this case. There are several different sorts of intermediary that could be relevant.

Our judgement is that it is too early to understand the eco-system of intermediaries that will be needed and thus premature to introduce accreditation schemes at this stage. That would stifle innovation at a time when it is most needed.

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.

One would hope that nothing more needs to be done regarding ensuring that public service providers respect privacy and security considerations - that should be embedded in their culture as well as existing legislation. The ways to ensure honouring of the commitment to open data were outlined above - lower barriers to release, gain benefit for stakeholders by encouraging meaningful comparable releases, review compliance.

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?

The correct way to constitute Board-level responsibilities should be up to the stakeholder organization. The goal is to meet legal mandates and Open Data ambitions; the mechanisms for best doing that will vary.

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

While the *carrot* side of the equation (showing the benefit, Government using its own data) is

the more powerful, some *stick* side will be needed. Include metrics on how well Open Data commitment are being met as part of regular review and evaluations.

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

NA

Meaningful Open Data

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

Data inventories are important to enable third parties to understand what data is, or could be, available; they are also needed to support discovery and reuse internally as well as externally. At present there are several related requirements for inventory and registration including Publication Schemes and Information Asset Registers, with new ones to be introduced as a result of the INSPIRE directive. Inventories to support open data should build on and complement these - not create yet another complex burden.

Linked Data approaches provide a suitable technical foundation for such inventories by supporting distributed creation (so that the inventories can be created and managed by the stakeholders directly but aggregated for discovery and sharing). It also allows the description and detail of the inventory to be adapted to the needs for each type of data set in the inventory. Again the Open Data leadership group will be needed to provide standards and coordination to ensure that the aggregate inventory is useful and fit for purpose.

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

The highest priority data is that whose value is multiplied because it unlocks the meaning and value in other data sets, enabling them to be combined and interpreted. Thus a priority should be given to reference data including identification schemes and coordinating vocabularies.

Beyond that it will not always be clear what data will be important so the presumption should be to include as much as possible in a lightweight way in the inventory to make discovery possible. However, there are some signals that categories of data are important and so need to be inventoried and opened, these include:

- data already in use in multiple places across Government, especially where this already leads to duplication and resource waste;
- data with clear relevance to policy decisions and consultation by providing context and evidence to support decision making;
- data which is frequently requested in FoI or open data requests and/or has been seen to stimulate innovation when released by other jurisdictions.

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

Wherever Government regulates or provides a service it should automatically provide and maintain reference data (identifiers, URIs, with core data associated) to enable Government and others to describe and reference the elements of the service and coordinate data about

it. The recent publication by Companies House of core reference data on companies is a good example of this. Reference data has to be maintained; users of the data need to be able to rely on it being updated and supported in the long term.

A coherent set of such reference data enables not just coordinated publication of data by Government but allows third party data holders to use the reference to link their own datasets. This provides the foundation for a *national information infrastructure*.

Beyond the reference data then we would expect routine collection and publication of performance and metric data which allows the costs, effectiveness and operation of services to be compared and analysed.

We assume that comprehensive collection and publication of local and national statistics giving a complete (economic, social, environmental, heath etc) view of current status and trends will continue.

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

Ultimately data that is never used by anyone, whether in Government or externally, should not be collected and maintained. However, it is challenging to separate data which will never be used from data which may be used eventually as needs emerge or new analysis technologies develop.

A first step is to identify actual use. The data inventories should be a useful tool in this. It should be possible for (external and Government) users to notify that they are reusing a dataset. This gives some basic information from which usage and evidence of benefits can be built. Datasets with no apparent internal or external use over some time can then be considered for review.

The data inventories should also help identify where data collection is being duplicated. Wherever authoritative data exists it should be reused from source rather than replicated or duplicated. More efficient and effective reuse of data within Government is a key benefit of greater adoption of Open Data principles, over and above the unlocking of external value added services.

Note that rather than think in terms of binary decisions over datasets it would be better to consider levels of investment. Retaining raw data that is collected anyway (c.f. weblogs or service call logs) is generally cheap. What is expensive is the refinement of that raw data to create high value datasets or the pro-active collection of new data through surveys and directed studies. In some cases it may be appropriate to continue to record raw data but reduce investment in refining and packaging it. External parties may be able to mine value out of apparently low utility data as technologies improve.

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?

Data must be fit for purpose. The appropriate quality metrics depend entirely on that purpose.

The most critical requirement is that the quality of the data released be defined and published along with the data so that (re-)users can decide to what uses the data can be safely put and to what extent they can rely on it. In many cases it may be appropriate to

release unpolished data, so as not to hold up publication, but the assumptions behind the data and the extent to which it can be relied on must be clearly communicated. This is especially important as we move from a world where Government mostly publishes very carefully controlled official statistics to one where less polished data is available. The infrastructure (in terms of agreed standards and descriptive terms as well as technical infrastructure such as use of Linked data) to enable this needs to be coordinated and actively invested in.

There are many dimensions to quality including use of quality control measures, accuracy, precision, timeliness, coherence, granularity, authority etc. For many data sets but especially for reference data we would argue that the critical dimensions to consider are:

- *sustained* - one off data sets are rarely useful other than as historic curiosities; to invest in using a data set or building value added services on top you need to know that the data will be maintained - the lack of commitment to maintenance is arguably one of the greatest reasons for duplication of data sets within Government;
- *authoritative, complete* - data which is complete by construction (e.g. provided directly by a regulator or statutory provider) should be prioritized and marked as such;
- *timeliness* - the delay between change occurring and it being reflected in the data needs to be defined; not all data sets need to be perfectly up to date but decisions that affect (for example) monthly spending cannot be based on data which has a 3 month lag;
- *defined status* - the provenance, quality control, accuracy and completeness of the data needs to be provided; unpolished data which clearly states its limitations may be very valuable for some third parties, unpolished data which could be mistaken for official statistics could inflict substantial harm if it is incorrectly used as basis for critical decision making.

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?

Do not think in terms of portals as places as where data is *held*.

The web, and the use of web principles and technologies for data management (Linked Data) have taught us that data can be published in a distributed fashion. Data should be held and managed close to source by the stakeholders who know about and care about this data. Web technologies can then be used to index and aggregate this data for ease of discovery and reuse. A portal should be regarded as simply one user-facing application built upon this data fabric, not a box in which all data must be placed.

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

The highest priority data is that whose value is multiplied because it unlocks the value in other data sets - enabling them to be combined and interpreted. Thus a priority should be given to reference data including identification schemes and coordinating vocabularies. Data, like information on the web, is subject to a network effect - the value of the network in providing context, comparisons and comprehensiveness outweigh the value of individual web pages or data sets. Priority should thus be given to data which helps create and enrich the network.

Secondly data that is known (from other data releases) to fuel innovation (for example, transport data), and/or is frequently requested, should be a priority. We cannot always predict the outcome of such innovation and it is better to respond to demand and see where that takes us than over-analyse the specific expected benefits.

Beyond that, then the consultation paper lays out a good framework for understanding the benefits of Open Data - transparency, empowering choice, improving productivity and outcome quality, social and economic growth. Important data sets are those which directly support one or more of these benefits.

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

This is a false dichotomy. In some areas we need data to be brought up to date and maintained but with the same depth (e.g. the Education linked data which is now useless because it is two years out of date), in some areas the data is simply not yet being published (e.g. timetable information), in others then finer grain detail to enable more localized use could be helpful.

To repeat our overall message - prioritize data that enables network effects, that gives meaning to both Government and external datasets, that can be widely reused, that fuels innovation.

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?

This is a critical part of the consultation.

Simply throwing data “over the wall” and hoping that people can make use of it will not deliver the full benefits back to Government and society and will not create a sustainable eco-system of intermediaries improving the cost/effectiveness of Government’s own use of data.

Government has to participate actively in the market place, becoming a consumer of the data services which are built on top of the open data releases. It should become best practice that a given data set be held and maintained authoritatively in one place, released as Open Data and different parts of Government reuse it at source rather than replicating or duplicating it. Examples abound of repeated data collection and duplication leading to waste (from myriad different central government reference data lists on local authorities to duplication of legislation data bases to duplication of spatial data collection due to the problems of OS licensing). To change this behaviour the data publishers need to commit to timely and sustained maintenance of their data sets (see comments on quality above) but often the data consumers will need new views and value added services on top of the data for it to be useful to them. By default they should create these not by building bespoke closed internal information solutions but by going to the open market with their requirements and let the market place innovate to provide solutions (whether refined, integrated versions of data feeds or visualizations, applications and decision support tools).

Other approaches to this stimulation have been shown to work less well.

Competitions to create new application ideas can create an initial burst of enthusiasm but without the promise of Government actively buying back some of these services there is insufficient funding flowing through the system to bring many services to an sustainable level. The net result of competitions without follow-on commitment is first a lot of interesting demonstrations and then a disappointment that none of these are turned into long term services.

Funding the *push* side of the equation through R&D funding has its place but Government is traditionally poor at spotting the right innovations to invest in. The *pull* model of Government being a source of requirements and funding as a customer gives much more direction and focus to the investment while allowing a freer open innovation to take place.

Note that it is important that such *pull* be open to a broad range of providers. Government, despite recent rhetoric, is not well adapted to purchasing of services from innovative SME providers. This must change if Government is to reap the benefits of innovation and opening data can generate.