Comments on the Public Consultation from the Cabinet office
Making Open Data Real: A Public Consultation

The comments below are based on my experiences of working in the public sector for approx 13 years. They do not relate to any specific public body and are my personal observations on the way public bodies work and having to manage data as part of my job.

Page 8 Paragraph 3.2
This initiative will not necessarily reduce the administrative burden on services. Currently a (probably large) number of authorities only collect data that they have to. Submitting data to government could mean more work for some public bodies.

Page 8 Paragraph 3.3
Does the public actually want to know about the internal workings of the government and public bodies? I think they are more interested in getting the services delivered than how it is done.

Page 11 Paragraph 4.7
This is true. Data collection for economic development is vital especially in our current economic climate. Data needs provenance/metadata which needs to be cited on any data that is published. It can be difficult/expensive for the public sector to get hold of/collect the core dataset and this is often why the data is not published. Often, data is not collected at a low enough level to make it meaningful for re-use and to target services/support in the right places.

Page 14 Paragraph 6.1
The possible implications of releasing 'raw' data before checking its quality are huge. This can provide skewed or erroneous results as data will not have been checked, for example, or may not have been input correctly. For example, the medical research on iron; for decades people thought spinach had a much higher iron content than anything else when in fact the decimal point had been erroneously put in the wrong place. If the public body has to release the data it is better that it is released once after it has been checked and validated than released twice because the first release contained errors. Paragraph 7.3 confirms this.

Page 20 Paragraph 7.5 & 7.6
There have been numerous stories in recent years about computer hackers who have gained access to personal information, bank account details even national security systems. Whilst making educational and medical records accessible is a good idea to make them available over the internet could turn into a hackers paradise. Making it possible for unscrupulous individuals to find out what medications are prescribed to an MP or their children, for example. Patients already have the right to view their medical records at the surgery.

If this does go ahead then people should also be given the opportunity to opt out. There are a lot of older people, for example, who do not have access to a

computer or the internet and they will not want their information being made available in this way. The public has a right to say what happens about their personal information.

If data is to be provided free to the public, and presumably local authorities will be expected to collect, maintain and report this data, then how is the cost of this expected to be born when local authorities are experiencing significant budget cuts and are concentrating on service delivery? Data collection and publication has traditionally been seen as a back office function. Collection of data can be expensive and thought needs to be given around which datasets are collected and whether they will provide value for money.

Metadata about datasets will need to be recorded and released with the data. If the government wants consistency then it will need to explain what datasets it wants collected and the method of collection otherwise consistency will be lost.

Page 22 Section 8

Key Area 1 – An Enhanced Right to Data
Review Copyright and Intellectual Property Rights legislation
If a business intends making money from public data the source should be acknowledged and money made from using this data should be shared with the organisations providing the data. It is unethical to allow others to make money without at least acknowledging the work and contribution made by others. This is the equivalent of plagiarism.

In view of the reduced budgets that public bodies are experiencing this additional requirement to proactively publish data on services will add further burden.

The Information Commissioner could take on the role of independent reviewing body.

Page 24 First bullet point

No the additional resource required to prepare datasets if the cost of FOI requests were increased, would not be proportionate. Most people underestimate the amount of time required to collect and manipulate data to get it into the required format.

Page 24 Fourth Bullet Point

Yes, a public provider should have the right to refuse to publish because of unreasonable cost. Like the Disability Discrimination Act, reasonableness should be applied. Yes, if the data requester is prepared to meet the cost and the data is not sensitive in some way then it should be provided. Yes, this could have an impact on the service provider delivering its core functions so delivering this data would have to be discussed with the requester, a time scale agreed and if the service provider is not able to provide the data themselves, then the work can be put out to an external organisation.

Page 24 Fifth bullet point

No this will not work. Modern systems should be ODBC compliant. It's usually the case that the IT dept does not want the job of writing another report. It is possible to get data out of any system, it is just more time consuming with some than others. Legacy systems are the worst because they have been written in a defunct language and programmes like Crystal reports can not be used.

Page 24-25 Paragraph 7

Changing procurement rules will not make any difference. It's the specification in each IT procurement project that needs changing to ensure that it is straightforward to extract data.

Page 25 Paragraph 8 and 8.7

Most systems purchased by local authorities are ODBC/open source systems and a lot of software providers would not be able to sell their software without this. Through the e-gov initiative the public sector has been steadily working towards publishing information online. Availability of resources is the usual reason why some organisations have been slower than others. Publishing data on the internet is as much about having the resources to get the data into the system as it is purchasing the system. Putting the data into the system is more expensive and time consuming than buying software.

The public sector does not have a track history of being an intelligent, demanding customer. Widespread change in the medium term is only achievable if the public sector is going to receive some financial support.

Page 26 Paragraph 8.8 and 8.9

This begs the question of what will happen to this lower quality data? If this poorer quality data will be used to compare with higher quality data then that will be like comparing apples with pears. If the data re-user knows the data quality is poor they will not use it. Set a standard from the start, give the public sector enough time to organise collection and place liability on the public body to provide quality data.

The proposed five star system is sensible. A target system similar to that used by GeoPlace for the National Land and Property Gazetteer (NLPG) may be useful.

Metadata needs to be included with any published data.

Page 27 Paragraph 8.10 fourth bullet point

Whoever wrote this sentence does not know what metadata is. Metadata provides the user with who, what, where, how, when of the dataset so the re-user can make a judgement about the quality and appropriateness. I suggest someone talks to the Location Council about the different international metadata standards and the work that is being done to comply with INSPIRE.

Page 28 Paragraph 8.11

Excellent suggestion, go for it!

Page 29 - Key Area 3 - Corporate and Personal Responsibility

Commitment to Open Data should be a corporate responsibility with a senior person taking overall responsibility for compliance. An experienced/trained person should deal with data protection and privacy, they need to understand all aspects of data management.

Responsibility would have to be added to job descriptions. Legislation will be required making clear levels of responsibility and punishment for not meeting these requirements, eg INSPIRE. This means that monitoring will be necessary.

If businesses and individuals are to use data produced by public providers then they must also be responsible for publishing accurate results and not manipulating data to gain a specific desired result. Like the science world carries out peer review, maybe something similar could be set up.

I suggest that a carrot approach would be more successful than sanctions although there needs to be some level of punishment for non compliance. A carrot will also mean that organisations will comply more quickly.

I suggest a Sector Transparency Board may be useful for all sectors where the government is planning the collection and submission data.

Page 31 - Key Area 4 - Meaningful Open Data

Monitoring would be required to ensure that public bodies are collecting data and publishing it. Maybe regular downloads to a national hub, e.g. data.gov where, like the NLPG, regular updates are required to meet agreed standards. Data from the providers should be held centrally so that the public knows where to go, e.g. data.gov.

Why not follow the INSPIRE model? Have a national hub where public bodies load their metadata. The re-user can look on the hub to locate the information they require and then use a hyperlink to go to the webpage where the data is held. Discovery level metadata could be recorded on data.gov.uk using one of the existing national or international standards.

Metadata could be uploaded each time there is a change to a dataset or a new one is published. This will make the data more meaningful.

Question 1 – talk to the Location Council, they are already working through this for geographical data.

Question 2 – What type of value? Monetary? Data has intellectual property rights but the proposals are to give the public's IPR away so it loses any monetary value. Create an inventory of the datasets that are already known and in use. Then get each Sector Transparency Board to look at their service area, define which datasets are required and prioritise them.

Question 4 – A significant proportion of performance data. For example, nurses going around hospitals with clip boards checking that all staff have

their sleeves rolled up. Public bodies spend their time and money collecting data for matters where they are measured, not necessarily on areas that are important locally.

Question 5 – Yes, data should always be high quality. It is far more costly to publish lower quality data, refine it and then republish. There is always an element of compiling and formatting even poor data before it released so this work would be duplicated if the same dataset is published twice or more. It will also mean that re-users are less likely to use the poor quality data and wait for the good quality data as they will have to compile the data or run their analysis twice too. This is not cost effective.

Defining quality will depend on the dataset in question. Polishing data implies tweeking or massaging the figures it to make it look good. This is not a good idea if the government wants quality.

Releasing public data will also mean that it will highlight areas of incompetence as well as good practice. The government should be asking what the information will be used for and that each time a re-user asks for data they have to sign a declaration about what they want the data for. People undervalue data and don't see it as important. It is possible that data can be used for criminal or immoral activity. I think the public has a right to know who has accessed what data and for what purpose. It's their data, paid for from the public purse. For example floor plans published by building control departments provide potential burglars with information about the level of security on windows and doors and the size of a house (if a person can afford a house of that size they are likely to fill it with valuable goods and/or drive an expensive car).

Page 33 - Key Area 5 – Government sets the example

Any underlying data behind advice and decisions should be published with and at the same time as the report/document. In principle publishing datasets along with the analysis is good idea. However it will mean that the public will question government decisions but will mean that policies and decisions have been made on matters of fact and will improve transparency. The government would need to ensure that the published datasets are accurate and have not been manipulated to gain a specific desired result. Inaccurate data would be embarrassing for the government and give the public another reason to mistrust politicians.

Prioritisation of datasets should be based on need. Each sector will have different priorities and whilst some steer from government will be required, each public body will have its own priorities so some flexibility about local priorities should be allowed.

Question 1 – I suggest a central national portal be used then the public and re-users will know where to go, ie data,gov. The public have difficulty understanding the difference between a County Council and a District Council. Different government departments having portals will only add to the

confusion and give the impression that the government is still trying to hide its data.

Question 2 – publish evidence of existing datasets behind regular statements first. Cut your teeth on what is familiar and tried and tested first. Take a phased approach to allow time for data creators and statisticians in public bodies to become familiar with what is required and the standards they must meet. Then gradually introduce new dataset requirements around new initiatives.

Question 3 – Improve what you've already got first before running off and spending money on collecting new stuff.

Page 35 - Key Area 6 – Innovation with Open Data

Improve the ICT infrastructure. The UK broadband service is poor and our current speeds are worse than some countries in Africa. Many countries have speeds greater than 100Mb/sec. Data can be resource hungry, you need a good infrastructure to get the best out of data. For example, small businesses need access to the internet so that their business can grow. Rural areas are highly dependent on small businesses for their economy. Rural areas have poor broadband speeds which get worse as the distance from the exchange increases. Poor infrastructure means fewer business start-ups, especially in rural areas. It also affects training and learning, people in rural areas are less likely to use the Open University, for example.

Page 36 – Question 1

Yes there is a role for government to stimulate innovation. Review legislation around copyright and intellectual property rights as this is currently cumbersome and confusing. Improve data.gov; a central place where people can go to see if data is available. Server sizes will probably need increasing.

Annex 1 Paragraph 1.18

There is fundamental problem with this statement. Often, it is not possible to put a patient's full medical records onto the system as older hand written information can be impossible to read. Even if this information can be read, it will take an extremely long time to get all medical records onto a computer system because of the quantities involved. So this needs to be seen as a long term project; it is simply not possible to get all this information accessible online in the medium term.

Even if these 60m records get put onto computer, the servers will need to be increased to cope with the demands on memory, broadband capacity will need to be increased to cope with the quantity of data being passed backwards and forwards and then there's the back up systems for disaster recovery. Will the government be funding these hardware costs?

Annex 2
Metadata needs to be published with the data.

Releasing data quickly is not a good idea as it will compromise quality. It also means a duplication of work and re-users will view it with suspicion. It may also impact on best practice.

Users should register for access to data. Current providers of free data require down loaders to register. This will allow the government to monitor who is accessing the data, ascertain which datasets are more popular, whether the uptake in the business community has been as anticipated, what sort of data is being used by the business community etc.

Those who use public data should also share their results and findings with the public/government as it may be useful for policy making, public health, economic development etc

Other General Comments
Data is like cars. You don't sell cars to the public in kit form for them to put together. Or they will end up with a few screws and bits left over. You provide them with the completed article which is safe to use and quality controlled. Car manufacturers only release their cars to be sold after they have had a level of quality control and testing done. Remember Toyota's problems last year. Most people do not know how to use data or interpret it. It is not taught in school and unless it is provided with an explanation it is unlikely to be meaningful to many people.

Anecdotal evidence is that public bodies concentrate their energies and funding on areas where they are monitored. Service areas which are not under scrutiny receive a lower priority.

Careful consideration needs to be given to third party intellectual property rights and copyright. Also thought needs to be given on who will use the data and for what purpose? What are the implications of making this data available? Whilst possibly stimulating the economy there can be adverse implications too. E.g. publishing crime data might affect house prices/desirability to live in a specific area. There is already anecdotal evidence that estate agents are advising clients to not report crime as it may affect the value of their house.

High quality data can be expensive and it takes time and money to make sure that data meets a certain standard. That means that there will be an up-front cost attached to getting things set up ready to meet these proposed requirements and afterwards a maintenance cost. This means that there will be a cost attached to public bodies providing this information and hence a cost to the public purse.

One thing to consider is that some information is still recorded on paper. For example, some councils still operate a part IT part paper based land charges search and doctors practices have part of the patients records on database

and part in paper. The cost of digitising this information would be phenomenally expensive, even if it can be digitised. Doctors' handwriting is notoriously difficult to read!

Private companies and academic bodies should also submit data. Utilities could provide data on how much electricity is going into the grid from wind farms and solar panel generation?

Truly open data should include the private sector making datasets they have collected available for use by the public sector too. For example, the Tesco ClubCard and the Nectar Card schemes collect terrabytes of information about peoples shopping habits. This information would be extremely useful for analysis of the economic market.

Rebecca Domek MSc FRGS MRICS CGeog(GIS)