Written evidence from the Open Data User Group (ODUG) [OD14]

Summary

1.1. The Open Data User Group (ODUG) is an independent advisory body appointed by the Chair of the Group and the Minister for the Cabinet Office. It receives administrative support from staff in the Cabinet Office, but speaks independently on behalf of the data community

1.2. This is a collective response from ODUG to the Public Administration Select Committee request for evidence on statistics and open data.

1.3. Individual members of ODUG, or their employing organizations, may make separate responses.

Our main observations are:

2.1 The release of Public Sector Information, which is not restricted on grounds of privacy or security, as Open Data, is the most effective way of maximising its utility and allowing the nation to achieve maximum benefit from the taxpayer funds used to collect the data originally.

2.2 In addition to re-use, which may generate additional economic activity, Open Data contributes to transparency enabling the public to hold the public sector to account for the effective delivery of public services. Where such data relates to the quality of public services it can often allow the public to exercise informed choice.

2.3 Making information available as Open Data often leads to improvements in quality as errors will be noticed and reported by more individuals and organisations.

2.4 Government policy relating to which data should be open, which should be closed, which should be traded and at what prices has been incoherent and has often appeared to be arbitrary and lacking in detailed accountability.

2.5 Despite the agreed policy of a 'presumption to publish' the release of public sector information as Open Data has often been inconsistently planned and is poorly regulated. There needs to be clear accountability where public officials are answerable to a single body or individual who will ensure that they are fulfilling their obligations to release Open Data.

2.6 All future systems to collect public sector data for the fulfilment of public services delivery should be required, from their inception, to build in an Open Data delivery mechanism which is publicly accountable and takes data security and privacy issues into consideration.

PASC's questions

1. Why is open data important?

1.1 Information has become the key economic resource of the 21st century. The economic efficiency and the competitiveness of nations will depend on the choices that are made over how data is collected and shared to maximise the opportunities of deriving information from it.

1.2 Private companies have a clear obligation to use data, in the same way as any other resource, to maximise the return to shareholders.

1.3 Some private companies achieve this by apparently making the data they collect, create or collate openly available. This is usually not the case as, often, users are only permitted to view the data, not to capture and reuse it. Making the data 'open' to view only is a mechanism to attract users to view advertising material that is placed alongside it, and data reuse is limited to prevent the collation of information by others.

1.4 Genuinely Open Data from the private and voluntary sectors is rare. Wikipedia and OpenstreetMap.org are notable exceptions; both are not-for-profit organisations that allow the relatively free re-use of their data.

1.5 In the age of broadband and superfast broadband the internet has become a ubiquitous channel of communication. Once initial data hosting costs are covered the marginal cost of sharing data is close to zero and allows the release of Open Data as a public good at a substantially lower cost than previous mechanisms for disseminating data which carried significant costs. It is this technological change that has created the Open Data opportunity.

1.6 Around the world many governments are coming to the conclusion that maximising the amount of Open Data available to citizens and businesses is the way to maximise the economic benefit that countries as a whole can derive from that data. This is particularly the case for Public Sector Information (PSI) which has already had to be collected in order to allow government departments or agencies to fulfil their Public Tasks. European regulations on the re-use of PSI strongly encourage the release of as much of it as possible as Open Data (subject to privacy or security constraints) in order to maximise its re-use and the economic activity related to its re-use.

2. Why does the Government need an open data strategy?

2.1. Government has had periodic major strategic reviews of the release and charging mechanisms for public sector information, in particular statistical information and core reference data. Perhaps the most significant ones have been:

2.2. The **Rayner Review of 1979**, which led indirectly to a Tradable Information policy being introduced by the Treasury which lives on in guidelines for charging for government information.

2.3. The Rayner Review created a situation where vital statistical information, and in particular Census Information was no longer freely available even to Government Departments, but had to be purchased through "Census Agencies", private sector brokers who paid a licence

fee for the exclusive right to re-sell government statistical information. This led to some government departments claiming that they were no longer able to afford to pay for the information necessary to carry out aspects of their Public Tasks.

2.4. **"Crown Copyright in the Information Age" published in 1998.** Rather than taking a strategic view of what data should be open and how it should be funded, this review divided agencies into those that were successfully recovering a significant proportion of their costs by trading in information, such as Ordnance Survey (OS), and should continue to do so, and those not recovering significant costs which were liberated to release their data without charge, such as the Office for National Statistics (ONS). The impact of this seemingly pragmatic decision was to create a difficult divide between the trading agencies and the Open Data agencies.

2.5. Addressing and geospatial data is of particular value to all aspects of our society. Its pseudo-commercialisation as tradable data underpins much of the complexity found in the public sector information landscape where many public sector bodies re-purchase data from one-another in an overly complex, inefficient system which re-cycles public money between publicly owned entities.

2.6. Further complexities in licensing and re-use of public sector information arise from this system particularly where so-called 'derived-data' restrictions are put in place by the data holders. For example Ordnance Survey place restrictions on both public sector and private sector organisations re-using *their own data,* once it has been combined in some way with OS data.

2.7. As an example, in preparing for the 2001 Census the ONS found itself unable to afford all the geographical data from the OS which would have been helpful in conducting the Census. Following the publication of 2001 census results in 2003, ONS found that it was unable to release the census output area boundaries (OAs)[1] as unrestricted Open Data, because of Ordnance Survey's commercial interests in that data.

2.8. **A consultation on 'Policy options for geographic information from Ordnance Survey' in 2009**[2] led to the decision to release some Ordnance Survey (OS) data as Open Data, and the Public Sector Mapping Agreement (PSMA) which makes OS data available for free to the public sector. This was a significant change of policy, though it did not appear to be part of an overall strategy and excluded the Royal Mail Postcode Address File (PAF). OS negotiated with a nominated "intelligent customer" as to which products should be released as Open Data. These included administrative and statistical boundaries in BoundaryLine and Postcode locations (though not the locations, or the text, of individual addresses). This was a significant step forward for the Open Data agenda, but did not deliver a coherent Open Data strategy.

[1] The OA is the lowest geographical level at which census estimates are provided.

[2] http://webarchive.nationalarchives.gov.uk/20120919132719/http://www.communities.gov.uk/publications/corporate/ordnancesurveyconsultation

2.9. The **Cabinet Office Open Data White paper and the subsequent Shakespeare Review** both make cogent and compelling cases for Open Data. However neither can be regarded as a strategy and, while they have led to the release of over 10,000 government data sets, these have not been released or prioritised according to any discernible strategic framework. Nor is there any regulatory mechanism to ensure ongoing publication, release or change update strategies for the data sets that have been released.

2.10. In order for the country to gain the maximum utility and economic advantage from Open Data a more strategic response is needed. There needs to be a single, permanent, locus in government for the regulation of Open Data release and for the dissemination of a coherent strategy which will outline: (i) what should be released; (ii) how it should be released; (iii) how frequently dynamic data sets should be updated or whether, by simply exposing dynamic data sets, users should track changes themselves; (iv) what metadata standards should be used (data.gov.uk delivers a quasi standard and the recent National Information Infrastructure guidelines for departments are an additional step) and; (v) what skills and funding are required to ensure that the Open Data strategy is sustainable.

2.11. In the short term the Cabinet Office has taken on this role enthusiastically and is achieving some good results. However the overall remit is spread across many public sector bodies including the Office for Public Sector Information, the Information Commissioner's Office, The Department for Business Innovation and Skills (BIS), data holders within individual departments and their agencies, Local Authorities and other public sector data holders. These bodies do not work effectively together which we suggest is as a result of (a) the historic complexity of the legislative and executive landscape; and (b) a lack of overall strategy and priorities for Open Data and Open Data standards.

3. What should the Government's aims be for the release of open data?
a. Are the Government's stated key outcomes in its Open Data Strategy the right ones?

3.1 The Government's stated aims for the release of Open Data and the stated key outcomes of the Open Data Strategy are entirely appropriate and laudable. We support them strongly.

3.2 However we cannot be sure that there is a coherent commitment to the Open Data Strategy from all government departments. The government appears to be resolutely focussed on relatively small revenue streams arising from some of the Trading Funds and a short term focus on raising one-off revenues through the privatisation of public assets, which include fundamental datasets. As a result both parts of BIS and the Shareholder Executive are pushing forward short-term policy decisions which undermine the fundamentals of the Open Data Strategy.

3.3 While the release of individual data sets is valuable, some data is essential to make other data sets meaningful. This is sometimes referred to as "Core Reference Data", items of data which will be used across many data sets as identifiers to show what a record relates to. This data needs to be maintained and disseminated from a single source. Examples are: addresses with postcodes and geo-coordinates; geographical codes for statistical or administrative areas; company registration numbers; VAT numbers; codes or standardised names for health or

educational establishments; NHS numbers; Social Security numbers, classifications for public administrations and their services.

3.4 Some personal identifiers are contentious and should not be released as Open Data except as part of a system that anonymises records and preserves and protects the privacy of individuals. Others are, or should be, matters of public record, these include Company Registration Numbers, VAT registrations.

3.5 The decision by BIS to allow Royal Mail to take the Postcode Address File (PAF) into private ownership as a commercial data set, and for Ordnance Survey to participate in the creation of GeoPlace LLP as a trading Value Added Reseller of PAF which intends to commercially exploit geographically referenced addresses appears to fly in the face of any Government commitment to Open Data.

3.6 The Office for National Statistics has consistently explained how essential a single National Address Register is for a wide range of statistical purposes from taking a reliable Census, through Census alternatives to the sampling and geographical aggregation of many other surveys.

3.7 For users of statistics it is very important to know what set of addresses are included in each statistical area so that other data can be matched reliably to data from ONS, or data can be submitted to ONS reliably. It is difficult to understand how this can be accomplished effectively if a National Address Register is not available as Open Data.

4. How can those engaged in open data, and those engaged in producing government statistics work effectively together to produce new data?

4.1. The availability of government datasets as part of the National Information Infrastructure, including a single underlying platform of Core Reference Data needs to be agreed, delivered and maintained.

4.2. Government analysts and statisticians should make the widest possible use of this data, to avoid the inefficiency associated with multiple dataset-capture and maintenance of essentially identical sets of public sector information.

4.3. The use of single data repositories will enable improved like-for-like comparison and peer review in the generation of management information and statistics. Open Data allows improved levels of analysis and innovation, and will also increase the quality of the underlying data (the more users of a given dataset the higher the number of issues which will be detected and can be rectified).

5. How can more statistics and administrative data of all kinds become more freely available?

5.1. By enforcing the current presumption to publish and making sure all new data collection and IT contracts: (i) do not duplicate what is already collected by others; (ii) are designed with the release of Open Data as an upfront requirement. This should apply to all public bodies i.e. Central and Local Government.

5.2. Sustainable funding is essential to ensure that public sector information is made available, and continues to be available as Open Data. This can be achieved by ensuring that the cost of exposing data on an open platform is included in the initial cost of any project requiring data (this is generally likely to be a marginal overhead on the basic data collection costs).

5.3. Where data is generated as a result of statutory registration, such as: Land Registration, registering to vote, being registered to pay Council Tax or Business Rates, registering a planning application or building regulations consent etc. the cost of registration should include an element used to make the data collected openly available.

5.4. A principle of charging those who cause Open Data to change, rather than those who seek to use it, should be adopted as government policy. Citizens already pay to register births, deaths, marriages, cars, companies, land, planning applications, building permits and many other transactions. Each of these cause official data to change. A fee charged to a person registering, or causing, a change in an official record, should include an element that pays for that data to be included and disseminated as Open Data. This is a much more efficient way of funding Open Data than closing the data and charging for its use. It requires fewer transactions, the sums involved are a small percentage of the fees being paid and collected anyway and the amount collected is directly proportional to the update effort so there is no need to speculate how many times data will be used to set a price. This is a fair, efficient and effective way of funding Open Data generated as part of a public task.

6. Is open data presented well and of adequate quality?

a. Are the formats of the data being published accessible, useable and understandable to the public?

b. What metadata is needed to make releases useful?

c. Who will use the data released?

6.1 (a) The default requirement for CSV format is adequate in the first instance. Moves to deliver Linked Data formats are welcome, but not essential. The provision of APIs is useful in some use cases, but it does not constitute a proper Open Data solution to deliver a service (ie: pre-analysed or aggregated results) without allowing for the bulk download of the underlying data.

6.2 (a) While releasing data in a useful format is helpful and desirable, the minimum requirement should be that data is first released in the format used to deliver the public task for which it was collected. This minimises delay and cost in data release and it is likely that the Open Data community, or value adding re-processors will quickly adapt the data into more usable formats. This is the "Raw data now!" principle proposed by Sir Tim Berners Lee, which states that the release of data should be the prime requirement, while improving it to make it more usable is secondary.

6.3 (a) Where possible data should make use of existing standards and reference common vocabularies to enable the comparison and combination of data, therefore enhancing its reuse. Open Data should comply with the Open Data Certificate https://certificates.theodi.org/.

6.4 (b) At a minimum metadata concerning the available data formats, size of dataset, provenance (data owner), latest update (date), frequency of release, data holder (contact details) and reference to any common vocabularies should be provided.

6.5 (b) A number of more complete metadata standards such as Dublin Core and UK Gemini exist and are useful. However the requirement to provide complete metadata to these standards is often a barrier to Open Data release. It is for this reason that we suggest that a very simple metadata record should be treated as essential, while full standard metadata records are regarded as nice to have and can be added after data release.

6.6 (c) Society will use the data. It should be freely usable by anyone; public sector, voluntary and not-for-profit sectors, private sector, researchers and individual citizens.

7. How successful has the Government's Open Data initiative been in changing behaviour in the Civil Service and wider public sector?

7.1. data.gov.uk has been successful in providing a focal point for Open Data and a mechanism for the release of datasets which are not currently traded. It has been successful in starting to change attitudes and behaviours in the public sector about the rationale for and opportunities to release public sector information as Open Data. However, the main focus is on meeting central government departments' information publishing needs. data.gov.uk needs to consider the data publishing requirements of the wider public sector, including local government and other public bodies. It is too early to say how widely this data is used, but the provision of a simple mechanism and a single location for data release and to locate available datasets is essential.

7.2. However the availability of this Open Data has brought to the fore the issue of Tradable Public Sector Information which is not only unavailable as Open Data, but if combined with, or used in the process of generating, other data sets prevents their onward release in a useful form, thereby breaking the Open Data model.

8. Which datasets are the most important?

a. What are the best examples of data being made open and resultant benefits to business or society?

8.1. By definition "Core Reference Data" sets are the most important, these are data sets that provide the identifiers needed to generate and combine other data sets to release value. Core Reference Data generally identify places, institutions and individuals. Unfortunately the term Core Reference Data is sometimes used loosely to describe the data sets that individual departments believe are most important to their operation, rather than the data sets that contain true "reference" or connectivity data.

8.2. The single most frequently cited example of a core reference data set is a National Address Register. Past decisions concerning the Postcode Address File and the publicly owned GeoPlace LLP company owned by Ordnance Survey and the Local Government Association have undermined, for the time being, the prospect of an Open National Address Register that many are calling for as essential to society and to delivering economic value from other Open Data.

8.3. An excellent example of a data set which has been made open after more than a decade of resistance from Royal Mail and Ordnance Survey is the ONS Postcode Directory, an open dataset which gives the administrative and statistical area codes for every postcode current and past. This data set allows organisations to create statistical information which can be compared to statistics from ONS, increasing the value of both. Previously this was an expensive data set because of the requirements of the GridLink agreement between Ordnance Survey, Royal Mail and ONS which was a serious constraint.

8.4. Important areas where the release of Open Data has made the most progress and delivered tangible results to-date are in Companies House and Land Registry data, and education, health and transport datasets.

9. How effective is the work being undertaken by the Cabinet Office to monitor the progress of Departments in publishing their agreed datasets?

9.1. The Cabinet Office has required departments to set out Open Data Strategies which are helpful to set out goals and milestones for the release of Open Data. Over 10,000 datasets are now available on data.go.uk. These are positive results. However, the mechanisms available to hold departments and other public sector bodies to account are weak, hampered by a disparate legislative framework with responsibilities spread across multiple bodies and the pace of delivery is relatively slow. There has been little focus over the years in setting out an overall strategy for the Open Data agenda, as exemplified in the recent Shakespeare Review. The programme also lacks any substantial economic analysis to determine which datasets have the potential to deliver the greatest value to the economy. If a wider strategic approach had been taken earlier in the programme there would be more robust evidence available to underpin the current debate that core reference data should be made more widely available as Open Data and should not be allowed to pass into private ownership.

9.2. Officials also struggle to engage with the wider community and businesses. Setting up the Open Data User Group (ODUG), as an independent voice for the data community, has proven to be a positive step in bridging the gap between Whitehall and the 'real-world'. The data request mechanism ODUG has set up on data.gov.uk is a demand led approach which allows the data community to evidence the main barriers to the use of public datasets, and to highlight the datasets which will deliver the greatest value if released as Open Data. This has enabled ODUG to produce business cases and evidence the need to prioritise certain datasets deemed to be of most value to the data community. It is essential that data.gov.uk continues to collect and create detailed evidence with the new presumption to publish agenda.

9.3. Another very positive step, following the Shakespeare Review, is the proposal for the National Information Infrastructure (NII) requiring Departments to create dataset inventories and to monitor the release of datasets from those inventories. However the initial basis for these inventories is rather ad-hoc and it is difficult to judge what is missing from those directories and for what reason. Also, the inventories do not include public sector datasets collated and held by local authorities, many of which are of high importance and economic value – such as the National Street Gazetteer, nor do they identify datasets which span the requirements of multiple departments and public sector organisations, where significant efficiencies would be derived by cross-public sector organisation working.

9.4. It is not helpful that agreement has not been reached on opening up some of the important datasets (in particular addressing and geospatial data) which are fundamental as an underlying platform of Core Reference Data for the National Information Infrastructure. These datasets are essentially (re-)purchased from the data holders exclusively for the Public Sector, including products made available under the Public Sector Mapping Agreement (PSMA) to a restricted set of public bodies, and the proposed Public Sector Licence for the Postcode Address File (PAF). In both these cases public funds are used to re-purchase data which was originally funded from the public purse for the delivery of a public task. Such agreements remove the pressure to open the relevant data sets more widely for maximum economic benefit. They also leave many publicly funded and/or publicly regulated bodies, such as housing associations, the utilities and charities unable to justify the additional expenditure necessary to use essential public sector information which has already been paid for twice by the taxpayer, but whose access is restricted.

September 2013