UK DATA ARCHIVE RESPONSE TO THE CABINET OFFICE CONSULTATION ON MAKING OPEN DATA REAL

The UK Data Archive is a department of the University of Essex. It is also the curator of the largest collection of digital data in the social sciences and humanities in the United Kingdom. It provides much of the infrastructure for many of the various data service infrastructure provided by the Economic and Social Research Council (ESRC) on behalf of the UK Research Community. Neither the services we provide nor the data we hold on behalf of others is confined to UK HE/FE.

This evidence is submitted by the UK Data Archive and the Economic and Social Data Service (a service funded by the ESRC and operated between the Universities of Essex and Manchester) and represents the independent views of this department and the service. It does not necessarily reflect the views of the Economic and Social Research Council (ESRC) or the Universities of Essex or Manchester. For further details please contact Matthew Woollard, Director UK Data Archive / Economic and Social Data Service (matthew@essex.ac.uk).

This response focuses on the aspects of the consultation questions which are relevant to data produced by government departments and public bodies, and therefore does not answer every question.

Glossary of key terms

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

The term data is used throughout the consultation document, but it is not clearly defined and appears to refer to a wide range of different types of material from formatted information to data about identifiable individuals. It is important to distinguish data sources and data outputs; aggregated data and microdata; statistics and raw data and data and metadata if the document is to be properly understood. A default expectation that all data is to be made available makes little sense without such a definition, as everything can be seen as data.

Much of the discussion relates to public service “administrative” data: there is a considerable difference between data on public services, and data about individuals collected by public servants. Making the former open is much less problematic than the latter, and if this consultation is to be profitable this distinction must be made clearly.

A possible definition for the purposes of this consultation is a unit of information which is gathered consistently for the purposes of informing public service policy, provision, planning and evaluation, at the level of detail appropriate for such purposes. This does not preclude the possibility that data may be a by-product of delivery, but does not restrict it solely to that category.

From the point of view of the UK Data Archive and our many thousand users, we might, for data which is produced by government departments and public bodies) distinguish between a) data created as a result of an administrative activity (within a government department); b) data created purely for the purposes of informing policy (e.g., a statistical survey, like the Census). However, if public bodies are taken to include HEIs then data which is created as a by-product of research should form an additional category. This final type of data is **not** discussed further in this document, as the RCUK has provided a response covering these types of data; however some (but not all) of the principles evinced here are applicable to these forms of data.

he definition of Public Data is not clear.

There are contradictions within the document relating to personal data. These are data which can be unambiguously related to an individual. Datasets containing anonymised

personal data have huge value in social science research and must not be excluded from this consultation.

The consultation document makes no reference to Records (as defined by the Public Records Act). Data may be defined as public records for which there is a long-term obligation to provide access to.

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

Without clear definitions of dataset in the question it is difficult to answer. From the point of view of the social science data user community we would suggest that all data created by government or a public service specifically in order to inform public policy or used to inform public policy, but created as a by-product of service delivery should be made available with the appropriate levels of information/security protection. Most data created and used in these circumstances have the potential to be made available under the Open Government Licence. Some do not, see below.

We believe strongly that quality thresholds are so imprecise as to preclude this as a factor for "release"; however, any data which is released should have a quality statement from the producers to prevent or reduce the possibility of misinterpretation. Ideally all data should be quality controlled before release but this may not prove practical without additional resources which are difficult to justify. Quality control is important as it is related to public trust in data produced by public services that help inform policy decisions.

Other potential measures may include an identifiable benefit to the public. However, this is hard to benchmark and negates part of the argument about innovative use of data. It is not always possible to know what benefit data may bring, so it is best to attempt to keep these tests as broad as possible. Much government-collected data collected for the primary purpose of informing public policy or collected as an outcome of service delivery can inform secondary research which in turn can influence policy.

Data which is disclosive and can harm individuals should not be made Open; data which has been provided on a confidential basis should be treated within standard ethical guidelines.

Data should not be excluded from public/research use because it is inherently disclosive. Anonymisation techniques are well-understood within the statistical community, and data may be made Open after such techniques have been applied. However, the research potential for some data can be significantly weakened if inappropriate anonymisation techniques are applied. It is essential to the research community in general that this consultation does not endanger access to datasets which are currently available under a licence or specific secure access conditions (e.g., the Secure Data Service is currently negotiating to provide controlled and secure access to Student Loan Company data) for scientific research purposes. Licensing arrangement need to remain in place to protect data subjects while not hindering research.

Data should not be excluded from Open status because it is perceived by the creators to be "too complex" or "too dirty" for public use. We recognise that there are problems here, however, we believe, it is the responsibility of the data creator (or any information provider like the UK Data Archive) to ensure that these issues are explained at a sensible level to preclude obvious misinterpretations, but we agree that it is not the role of a government department to teach data analysis to the public! However, the opposite holds true also; if a government department produces low quality data for public consumption this may (and perhaps should) be taken as an indicator that that department is making decisions based on irrelevant data, and should lose the trust of the public in that area.

Some data may need to be embargoed to prevent use. Embargos should never usually be for more than clearly defined set period; however, the information about the data should be

released and the date of release should be publicised. Where possible the embargo should be explained. This period should not be considered to be a period for enhancing the data to meet whatever standards are adopted for Open Data publication. We also believe that different requirements **may** be necessary for data produced within a research environment.

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 acceptable to recover data publication charges where the request is non-standard, but if there are a large number of users demanding a dataset and it already exists, that represents a strong case that it should be Open because it is a good indication of likely utility and hence realisation of six opportunities of open data outlined in consultation paper.

The Economic and Social Data Service (ESDS) currently provides free (at the point of use) access to data which it holds on behalf of government departments and other public bodies. Some data are only available to particular communities owing to restrictions imposed by data licences or other access restrictions. Charges are applied if data are used for commercial purposes in accordance to our charging policy.

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?

As noted above all data which is used to inform public policy or the activities of a public body should be within scope, so essentially all organisations which are subject to FoI should be subject to any policy resulting from this consultation.

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

The most important mechanism is to inculcate an attitude towards making data available based on the arguments provided in the briefing paper for this consultation. Some evidence may need to be provided to back up the claims made for improving public choice or economic growth, for example. Overall however data sharing should be enrooted and widely supported.

Freedom of Information legislation may provide the legislative framework under which Open Data can be provided.

An Enhanced Right to Data

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

Simple recognition (and reward) may suffice, but building this expectation into Freedom of Information legislation should provide the “stick” where necessary.

Further research into the benefits of Open Data, and promoting sensible case studies showing both the obvious and the less obvious impacts will help.

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?

Yes, in general. There may be some exceptions where the Information Commissioner does not yet have the ability to recognise the importance for research purposes.

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

In general yes, but disclosure control of individual level microdata (e.g., the census) is not perfect, and must be combined with principles of safe data use if data utility is not reduced. The ONS Draft Research Data Access Policy rehearses the classification of microdata by impact level and then maps classification levels to access arrangements. The Open Data agenda can be seen, in essence, to deal with IL0 data. As noted above the creation of an Open dataset which has been anonymised must not be seen as a replacement for the release under special licence conditions to a disclosive dataset; it should be seen as an alternative.

The UK legal framework for data protection and confidentiality is ambiguous. As a result, the law is sometimes unnecessarily interpreted in a way that restricts openness and consequently the potential to benefit society and the economy. It can be argued that the protection in the Data Protection Act given to "personal information" (as distinct from "non-personal") is unhelpful, and that the better focus would be on protecting information that is entrusted "in confidence."

We suggest that the principle of minimisation of risk is preferable to the complete removal of risk. Some people may have the opportunity and ability to identify people by linking Open data with other microdata but they have neither motive nor willingness to do so.

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?

We understand the considerable costs of providing access to data, especially over the long term and applying industry standards to operate such services. The resource implications might be reduced by centralising access to these data, just as is currently provided by the data.gov.uk portal. However, this is currently an immature service which provides little in the way of coordination of data provision. If the framers of the consultation document believe the claims for the economic and social benefits of Open Data, then the benefits for "UK plc" will outweigh the costs many times over.

We believe that it is legitimate to use public funds to support the costs of storing and sharing publicly funded data, and that the costs of preparing data for storage and depositing it in an appropriate repository. These should be covered as part of the project costs, and should be budgeted for when before the data is being collected. There is growing evidence to suggest that the "additional" costs for the provision of high quality, well documented data are mostly removed if good quality data management principles are invoked at the beginning of the data life cycle.

It is also worth noting that the cost of "closure" of data may end up being more resource intensive, especially if the powers given to the public under FoI are more widely invoked.

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

By writing them into the contracts.

Setting Open Data standards

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

Contractually oblige any supplier to meet those standards, or legislate: the DDA (and, later, the Equality Act) has done wonders to ensure accessibility to government websites.

However, the "standards" proposed in the consultation document are not actually standards in any meaningful way. There are plenty of examples of good practice in data dissemination which could inform meaningful standards.

It is possible that government departments and other public bodies will eventually be "ashamed" into providing Open Data in formats which are not acceptable for use.

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

Possibly, but we do not see the collection of user experience across public services as being particularly germane to the needs of Open Data.

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

Absolutely, information intermediaries will become a significant force in interpreting "open government data" for clients and there will need to be regulation and accreditation. For certain types of scientific data there are existing schemes for the certification of trusted digital repositories: ISO 16363 on the "Audit and certification of trustworthy digital repositories" is likely to be too rigid for the purposes discussed here, but it could easily provide a reference document for the types of activities discussed here. We would suggest the Data Seal of Approval guidelines (http://www.datasealofapproval.org) should be consulted to provide more a research-based framework which could be adapted for these information intermediaries.

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?

Ensure that there are sufficient controls in place. The UK Data Archive, for example, already acts as an agent for public service providers in the dissemination of microdata produced by a number of government departments. In over 40 years worth of activity we have maintained a balance between the commitment to ensuring data is available for research purposes while ensuring that access is controlled in accordance with the risk/impact level of the data in the disclosure of data.

The UK Data Archive, on behalf of the Economic and Social Research Council already runs the Secure Data Service – a unique service in the UK with a philosophy of enabling secure and safe and yet remote access to sensitive data produced by public and private sector organisations. The main issue in providing access to 'controlled data' is to reduce risk to the lowest possible level without completely removing it.

If adequate guidelines are in place to ensure that data which has a potential disclosure risk are not inadvertently released to the public, this should be straightforward.

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 are properly protected and that privacy issues are met?

Potentially, but an independent body might provide uniform safeguards, cf. the Caldicott Guardian.

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

Quite possibly, but this would need to be tempered by the other legal and ethical pressures which currently exist, including the Data Protection Act.

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

No comment.

Meaningful Open Data

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

It is not clear from the consultation document whether a data inventory is supposed to mean a complete list of data which has been collected and *could* be made available, or a list of data which has been collected and will be/has been made available.

If the former it is possible that they could provide a mechanism for making a choice about precisely what data is made available. They would also provide a excellent mechanism for acknowledging data what data is collected by public services in a coherent way. If coordinated they could provide useful information on gaps in data and thus inform government departments and public bodies in what additional information may need to be collected to inform public policy. They may also have the potential to inform departments where there are potential overlaps in data collection, demonstrating redundancy and thus leading to efficiencies. Data Inventories might also include information on data use to inform public services collecting data on their actual use and hence rational for further collection.

It is important that the appropriate contextual information is included within any metadata provided within these inventories, so that potential users can assess the relevance and importance of these data for the use which they may wish to put them to.

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

This question suggests that there is some lack of clarity of what an inventory is. An inventory should be a complete list of all data which has been collected, rather than a list of only those data which are to made available.

User feedback and requests for data could provide a useful tool to establish perceived value.

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

As noted above.

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

No comment.

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?

This consultation should define quality more clearly. Quality may refer to a) the inherent quality of the data collection, i.e., collected without recourse to an appropriate methodological framework, e.g., poor sampling frame; b) the quality of the data itself, i.e., unusable because of a lack of consistency of key variables, e.g., the multiple spellings of Accenture; c) the quality of the metadata, i.e., poorly defined and labelled variables which could be confusing to a non-specialist; d) the use of non-standard groupings for variables, e.g., grouping ages of people inconsistently, e.g., 41-45 rather than 40-44 which would allow data linkage.

Some of these quality issues should be controlled for during their creation, but there is little use in releasing data of such low-quality that it cannot be taken as statistically significant for the uses to which it may be put. If low-quality data are released it may reduce trust by the government department or public body. However, in the interests of transparency it may be valid to make these data available so that the department or government body improves the data it collects and uses to inform policy, service delivery, planning and review.

Where 'polishing' takes place prior to publication this should be clearly indicated when the data are made available so that those using the data are fully aware of its provenance. One of the key transparency issues surrounding the release of data is that it allows people to understand on what basis decisions are made in government. If the data released provides different evidence, then it will reduce trust.

Whatever the quality of the data, the relevant metadata, including machine-readable metadata needed for automated semantic mining, should also be published to ensure the resources can be used in context.

Open data should be released in usable formats which allow re-use and re-purposing and linking with other data.

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?

Pragmatically both solutions may need to be followed, especially during the transition process to full open data. A further solution may be the creation of a "meta-portal", possibly based on a RDF triple-store, which aggregates the metadata to enhance resource discovery, allowing individual departments to manage their own data stores.

The provision of a portal should be seen to provide three key activities: 1) the store for data and its related metadata and other contextual information, which is managed and backed-up; 2) the resource-discovery mechanism which allows users to identify and request access to data; 3) a centralised standards-based information bank which allows producers to understand the file formats and provenance information which are required for the data.

It is again worth noting that the ESRC already funds the Economic and Social Data Service and the Secure Data Service which provide access to some to government and public sector data, so not all data may need to be stored in one central store; data of higher value for academic research may be hosted externally to this proposed portal.

What is not addressed in the consultation is any thought of the length of access to these data. There is no program for the long-term preservation of these data, and their long-term usability. This may be best achieved by the provision of additional infrastructure to allow for this, which may be best provided by The National Archives.

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

User groups may provide the most useful information here. However, the potential to inform research must be seen as an important factor. Further, the cost of data production may also provide a proxy for prioritisation. If a dataset costs £5m to create, it is likely (not always though) to have more value than a dataset which costs £5k.

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

This will clearly depend on the data; both are relevant in different circumstances.

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?

There are opportunities to stimulate innovation, but they are dependent on the right data being available to address important questions. Stimulating the open data initiative itself will produce innovation, but offering financial incentives or prizes may help in the short term.

In certain areas (e.g. social science, medical) there are already clear examples of social policy being determined by availability of open research outputs. Making public data (on e.g. health, education) available to social and medical research can be the starting point for new investigations, which can lead to improved public policy that can have wider benefits for society. Government's initiative on making public data public is a major step to provide a platform for free re-use of national and local data to facilitate the advancement of public services across the country.