Making Open Data Real:

A Public Consultation

Submission by Prospect to the Cabinet Office

October 2011

www.prospect.org.uk

Introduction

1. Prospect is an independent trade union representing 120,000 professional, managerial, technical and scientific staff across the private and public sectors. We are significantly the largest of two trades unions representing staff at Ordnance Survey and the Met Office. Our membership spans all grades and functions and includes the majority of the overall staffing. We are fortunate in being able to draw on our members' expertise in framing our response to this consultation.

2. Prospect's response to particular questions in the consultation document is set out in the paragraphs below.

Section 1: Glossary of key terms

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

Broadly speaking the definitions go far enough though it would be useful to have definitions of what is meant by 'structured' and 'unstructured' in relation to datasets and at what point a dataset is determined to have become 'Information'. For both Ordnance Survey and the Met Office data is not typically collected as a by-product of delivery but rather as a fundamental stage in the delivery of the organisations' key tasks which involve production of high quality, up-to-date information. It would be useful to have an additional definition for data which it is agreed will be made available but at a cost.

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

The initial tests should be around affordability, content of dataset, richness of dataset and whether the dataset is of use. The question then is who makes that judgement taking into account further tests around accessibility of the dataset, cost of provision as against benefit of use.

Both the Met office and Ordnance Survey both collect and produce data. Many, if not all, of the Ordnance Survey 'datasets' would be of little value for release and re-use by others unless enriched with sufficient meaning and context to render them 'Information' as opposed to simply 'Datasets'. The cost of converting the dataset to information is significant.

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?

This is a critical point, what measures need to be put in place to assess value against costs? Who creates the case and who judges the case? Clearly a business model needs to be sustainable against the price a customer is able, or is willing, or has to pay. This is in contrast to a model in which data is provided free and the cost being picked up by taxpayer through direct subsidy rather than it being bought. For example, the taxpayer bears the costs of quality control before Met Office data is released. There may be a hybrid approach, where users pay part of the cost and taxpayers also pay part, but this does raise a question as to whether others should be able to generate profits from data that is gathered and produced at taxpayers' expense. We accept that the Government's view on a data utility model is correct at present.

Requestors should be required to pay for public services data where the data is not collected routinely as a by-product of delivery and / or where data collected must additionally be enriched or put into more accessible formats to enable it to be of use. The costs paid should at least reflect the costs incurred by additional activities relating to enrichment, format accessibility and means of distribution / publication.

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?**

It would seem sensible to approach this question by way of first undertaking an audit along the following lines?

- Which organisations currently generate or could generate datasets as part of their provision of public services?
- What datasets or information do they currently produce or could they produce? Of these, which would be constrained by issues of privacy of personal information, national security, or commercial sensitivity, and to what extent?
- How should each of these datasets be categorised: i.e. raw data, value added data, information; and by what criteria should these distinctions be judged?
- What would be the costs of collection, collation and publication of these datasets in an accessible format? [1]
- What is the perceived value of publication and what level of interest is there in publication of particular datasets or information?

For those organisations which already produce data this would establish the cost, value, feasibility and practicality of publishing existing data to a particular level of accessibility together with determining the same thresholds in relation to public service providers which don't currently generate data but which could if there was an appetite for and value in them doing so.

Furthermore, the question should be asked as to whether data is being published primarily for purposes of transparency simply to inform, or for wider re-use, enrichment and commercial applications?

Criteria could be established to determine the use to which published information will or could be put which, in combination with the factors established by the above audit process, could then determine whether something should be published and, if so, in which format and whether there should be an applicable charge for provision.

Even where data is collected routinely as a by-produce of delivery of a public service, there is still always a cost associated with the creation of the data, more so when data must be enriched, or adapted prior to publication or where its' collection, collation and publication represents additional activity for an organisation.

Should others be enabled to generate profits out of data gathered and published at taxpayers' expense? Whilst provision of such data might generate economic activity and jobs, what guarantees are there that this would generate returns to the Treasury to off-set the cost of provision of datasets, particularly where those wishing to utilise data may be large multinational corporations who may then subsequently impose their own intellectual property rights upon enriched data, and where taxes due on revenues generated may be funnelled instead to overseas Treasuries? Managing Public Money guidelines may be of use in determining appropriate mechanisms for the extent to which organisations should fund provision of free data.

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

The provision of applicable datasets / information in a timely fashion and to an agreed standard could form part of any applicable public service providers Key Performance Indicators, with Board-level responsibility allocated for meeting targets set in this regard.

---

[1] in line with the guidelines developed by Sir Tim Berners-Lee: 'Five Star Ratings for Open Data', see http://www.w3.org/DesignIssues/LinkedData.html

Section 8: Policy Challenge questions

An Enhanced Right to Data

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

Subject to the criteria established above concerning what it is feasible and cost-effective to publish and in what format, a stronger presumption in favour of publication might be established by proposing that publication of data (as distinct from enriched 'Information') or at least the allowing of data to be published should be the default position, save for where issues of cost, privacy, national security, accessibility etc. indicate otherwise.

An Open Government agenda for greater transparency regarding the internal processes and performance of bodies within scope should in and of itself drive an enhanced right to data. If it is recognised and accepted that raw data held is data effectively 'owned by the electorate' then there will be a drive towards greater publication of data, particularly where the public feel that data is being unreasonably withheld.

7. 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?

It is difficult to say whether such an independent body would be the most effective option for safeguarding rights to access and data though it is clearly an option, and Prospect can see the benefits of the Information Commissioner producing clear guidelines or rules. Augmenting the current role and responsibilities of the Information Commissioner might be a more timely and cost-effective mechanism than looking to implement a new structure to undertake much of the same remit.

Other options might include expanding the number and duties of existing roles within bodies in scope surrounding data protection issues so that each public service provider would be responsible for ensuring delivery of rights to access and data within the confines of what is agreed it is relevant, feasible, and cost-effective for each provider to give access to. Performance measures surrounding delivery of enhanced rights to access and data could be linked to departmental KPIs as mentioned above.

Even if in-house departmental arrangements are adopted there is still likely a need to have a body to which appeals may be made in the event of refusal or failure to provide access to data which is deemed suitable for publication at some level / format.

We feel there is also a need to have a body that has oversight of what information is provided and to what use it is put; this to safeguard against data being provided at taxpayer expense free of charge to individuals or corporations, who subsequently turn this into profit-vehicles from which no benefit is derived by way of returns to the Treasury and the taxpayer.

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

Redacting can be used if the data is personally checked, line by line, to remove personal information. However we know this is not always the case and would in all likelihood add significant costs to the preparation of data, even where data is collected as a by-product of service delivery.

Should any data that relates to personal information be allowed in open-data and if not, what would the implications of this be for the quality and value of datasets being made accessible?

The existing safeguards appear to be acceptable to the public though this may in part be due to the extent to which they are aware or otherwise of how well existing safeguards actually work.

9. 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?

Where there are resource implications in making data accessible, the default provision could be such that publication is allowable within defined parameters, but that data will not be automatically published (and the costs of doing so incurred) unless specific requests are made for access. There is statement in the consultation document that 'The presumption is that data about public services will be Open Data. This could suggest that Open Data applies more widely to the activities of the organisation, for example including internal cost data, rather than simply to the scientific or geographic data it produces. Clarification on this point would be helpful.

If demand is for more than a basic level of access – perhaps in line with one or two stars under the Tim Berners-Lee proposed grading - then the cost of provision needs to be assessed. If the provider refuses to deliver then there needs to be an appeal process to ascertain if the data should be in the public domain even though there may be related cost. Paying the additional cost should not be a burden on the provider if the data has to be created. The cost may be picked up by the appellant, which may be a way of transferring the financial burden but may also be a tool used by the provider to limit open data.

The release of Open Data needs to have rules that allow access but the cost of access needs also to be a factor. Categorising datasets as 'suitable' for publication (subject to applicable licensing conditions perhaps?) might constitute a reasonable way forward wherein datasets need not be published unless / until specifically requested, and such that additional costs (actual or notional) are attributable to the individual/organisation making the request.

Having an expanded range of data which may be made public linked to licensing conditions which can accurately reflect the costs of publication to a given standard or format, the nature of the individual or organisation making the request and the use to which it is intended to be put might achieve a good level of balance between access and cost of provision and where these costs might best be attributed.

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

It could be made a condition of any ICT contracts that all hardware and software systems commissioned are compatible with existing systems and protocols for publication and distribution of relevant datasets. Some forms of information may require different standards of accessibility particularly where specialist software / hardware requirements exist.

An audit of the form described above would help in identifying what types of dataset / information exist and what the relevant requirements are for making these accessible. It should be straightforward from there to ensure that all new systems have sufficient 'backwards compatibility' to talk to existing systems or to employ sufficient resources to ensure that all applicable datasets are accessible in common formats.

Setting Open Data standards

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

An audit should be undertaken to determine what is relevant, feasible and cost-effective to publish, and in what format / level of accessibility (the Berners-Lee framework of star ratings could be used). Having determined what acceptable standards should be relative to quality of data, currency of data, and the extent to which data can be re-used rather than simply accessed, consideration should be given to making delivery against these standards part of a provider's KPIs. This provides an in-built incentive for each provider to deliver high performance but may have considerable cost implications, particularly if existing datasets are to be brought up to specification alongside ensuring all new datasets are compliant. Transparency and publication of providers' performance against established standards could itself form an open data dataset.

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

Yes. Government should retain responsibility for provision of access to government-owned, taxpayer-funded, datasets and information, and for monitoring how well this is delivered.

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

The addition of information intermediaries would likely only serve to introduce an unnecessary level of administration in the provision of data it is agreed should be made open. The individual providers producing the data are likely to be far better placed to understand the data they are responsible for, and to be responsible for ensuring it is made accessible in line with whatever protocols are agreed.

There could be an accreditation scheme for providers themselves to become accredited in the provision of their open data which might serve as a further incentive toward delivery of agreed open data provision.

To enable ease of access to data, a centrally-based portal could be established to which individual departments and providers could upload their open datasets for download by requestors.

Corporate and personal responsibility

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

Consideration should be given to making delivery of access to Open Data part of an organisation's Key Performance Indicators and/or including reference to delivery of data to the agreed level applicable to that provider part of their public task.

15. 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?

Delivering appropriate access to applicable datasets could, as stated above, be made part of an organisation's Key Performance Indicators. This could be expanded to ensure that named directors have personal responsibility for delivery against established standards and this could be made part of the package of performance indicators driving individual reward and remuneration for such named directors. It would seem sensible and appropriate that the same person/people be responsible for ensuring privacy and protection of personal data.

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

Setting KPI targets against delivery of open data should be sufficient, though an appeals mechanism to address issues of disagreement over access to or quality / usability of data might be useful. However, this would have to be set against agreed standards of what is to be made open and in what manner.

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

The current proposals for Sector Transparency Boards for Health, Education, Transport, Crime and Justice and Welfare seem sufficient.

Meaningful Open Data

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

data.gov.uk already exists and appears to provide a reasonably user-friendly portal through which to access Open Data. If providers are encouraged to establish individual inventories to house their particular datasets then these should at least all be linked or searchable through a single portal such as data.gov.uk and work to commonly agreed standards, otherwise requestors will first have to determine where to look for what they hope or think may be available.

The functionality of data.gov.uk could be enhanced to include email subscription or RSS feed type information services to inform subscribers when new datasets are added to categories they have expressed interest in receiving updates on.

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

An audit of the type described above would provide information as to what data exists; whether it should be included under an Open Data category; what format it is currently in; what level of re-formatting would be required to make it useful and the cost of doing so. This would seem a sensible juncture at which to convene further consultation amongst potential users as to what level of priority should be given to preparing and publishing particular datasets. It would be wasteful to spend time and resources preparing and publishing data for the sake of compliance with a standard where there is no expressed interest from potential users to have access to such data.

Establishing value is a much harder question, not least if it is unknown at the point of request or release what purpose the requestor has in mind when using the data. It is relatively straightforward to determine a cost of publication, at least in terms of time taken to collect, collate and prepare data for release, but this only gives a partial view as to the value of the data in question.

Broader questions would need to be addressed such as: the extent to which release of data without licensing conditions attached on the use to which it may be put impacts on the ability of the provider in question to continue to operate. This is particularly true of organisations such as Ordnance Survey where free release of the totality of their datasets would effectively prevent them being able to continue to collect high quality geographic information without considerable direct government funding to make up shortfalls through lost revenue.

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

Those areas which already have Sector Transparency Boards established though care should be taken to avoid Open Data creating a culture wherein too great an emphasis is given to the creation and publication of Open Data as distinct from the performance of the core public tasks for these areas.

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

Prospect is not aware of data which is collected unnecessarily but neither are we aware of the full range of datasets which are collected across public service providers.

22. 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?

Government should aspire to publishing data that is of a high quality though this needs to be balanced against the costs of doing so. Historical datasets will be harder to publish in more accessible formats than datasets collected in future according to agreed common standards and frameworks.

Quality can be judged according to a number of criteria; including but not limited to: completeness; accuracy; currency, as well as with reference to its' accessibility and flexibility in use as per the Berners-Lee framework of star ratings.

Whether public service providers should polish their data depends on a number of factors. Is the dataset of practical use without being polished or does the dataset require such polish in order to be meaningful to requestors? What are the cost implications involved in adding this polish and are these additional to the providers' normal day to day activities? If adding polish is required to render datasets meaningful, and this involves additional resource implications for the provider then this should inform whether the dataset should be published at all, or whether if published it should be charged for at a level which at least meets the cost of its production in that form.

Government sets the example

23. 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?

A single central depository of all Open Data might quickly become unfeasibly large and difficult to maintain despite offering the attraction of a 'one-stop-shop' for all Open Data. Individual departmental depositories would be more manageable, but it would be awkward to keep track of what was held where such that accessibility and transparency could be diminished.

A sensible solution may be to provide a single central portal which serves as a catalogue of all available Open Data, the storage and retrieval of which is managed through individual departmental systems. Departments and other public service providers could be held responsible for maintaining their individual systems, whilst ensuring that the central portal is updated as to available content.

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

Once it is established what datasets exist and which are suitable for publication, priority might be established by a number of factors including but not limited to: cost of publication (may vary depending on standard to which it is being prepared or 'polished'); value; to what use is the dataset likely to be put and what are the perceived benefits derived from these; appetite amongst potential users for a particular dataset to be made available.

Datasets which are able to be published with little or no additional resource implications could be published as a matter of course. Other datasets, for which there may be an appetite for release but for which there are significant resource or lost revenue implications despite perceived benefits deriving from their publication, could be given a higher priority provided the resource or revenue-loss implications are addressed. Such datasets are likely to fall more into the category of 'polished data' or 'Information' for which licensing or purchase arrangements are more likely to apply than publication free for all use and re-use.

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

For which is there greater demand? The priority should be determined by factors such as those discussed above. Where existing data can be published at a more detailed level without incurring significant additional costs or revenue-loss implications then there is little reason not to do this, but

whether this is more or less important than broadening the overall range of published data will largely depend on what appetite exists for better detail or broader range.

Innovation with Open Data

26. 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?

Publication of a broader range of datasets and to a higher quality will itself, if coupled with sufficient awareness-raising as to what is available for use and re-use, serve to stimulate innovation. Whether the Government should be more actively stimulating innovation would depend on what the perceived benefits versus costs of doing so are. There are good arguments in favour of maintaining and developing sufficient skills and knowledge within the public service providers themselves to enable public and civil servants to be at the forefront of innovation utilising the data they are creating in the course of delivery of their public task.