LinkedGov's submission to the Open Data Consultation

Making Open Data Real: A Public Consultation

#opendata

#openuk

Autumn 2011, Submitted to the Cabinet Office

Prepared by
Hadley Beeman.

Glyn Wintle

About LinkedGov

LinkedGov is a project to collaboratively clean, link, reformat and make useable all available UK public data. We are undertaking all this work with combined approaches of:

- Building an online factory to effectively complete the technical processes
- Creating games and fun activities out of the tasks which require human attention, making it simple to distribute the work.
- Highlighting the utility of the data and demystifying it for much of the public sector, encouraging them to help further with the publication and cleaning of the data.

The resulting clean, linked, usable data should be available for any purpose.

LinkedGov is a not-for-profit initiative that grew from the community. LinkedGov is sponsored by IC tomorrow, a programme run by the UK Government's Technology Strategy Board.

Table of contents

1. The aspirations for an Open Data policy

LinkedGov recommends these aspirations:

1. **Government and public sector bodies should publish all non-personally identifiable data as open data, for free and unrestricted reuse.**
2. **The data should be as up-to-date, accessible, versatile, available and understandable as possible.**
3. **Strong communication pathways between data publishers and data users should be established and maintained to help the public sector meet the needs of data users.**

We put these forward as aspirations, and recognise that certain practical considerations may get in the way of those principles being realised. As a result, in this document we have set forward our recommendations for accommodating those considerations for the smallest impact on the Government's goals.

1.1 Aspiration 1. Government and public sector bodies should publish all non-personally identifiable data as open data, for free and unrestricted reuse.

We would like the public sector to consider all non-personally identifiable data for publication, not just data on the performance of specific public services.

1.1.1 Types of data and why they matter

We often refer to the four broad types of public data outlined by Paul Clarke:

The four types of public data

1. Historical data

What's happened in the past: *how organisations and people have performed – what's been said in meetings – what's been spent – where the pollution has been – how children performed in tests…*

2. Planning data

What's projected to happen, or will shape what will happen: *this and next year's budget – legislation in progress – consultations – proposed housing developments – manifestos…*

3. Infrastructural data

The building blocks of useful services. Boring stuff, doesn't change that often, but when it does, it needs to be swiftly and accurately updated: *postcodes – boundaries – base maps – contact directories – opening hours – organisation structures – "find my nearest…"*

4. Operational data

The real-time stuff; what's happening NOW: *where's my train/bus? – crime in progress – emergency information – school closures – traffic reports – happening in your area*

today… [1]

Historical data is generally the most familiar: it is collected by public sector bodies and used to report their progress to other parts of government, develop business cases for funding, complete audits, and notify citizens (often through the press) of trends and changes. These processes have been in place largely since our public sector began using computers, which means that the data is reasonably easy to publish.

We equate historical data with the public services information the Prime Minister committed to publish in his letter of July 2011. It is intended to provide citizens with better information to guide their interactions with public services (for example, school spending data, school performance data, pupil cohort data and Ofsted judgements should help parents choose a good school for their child).

Though useful, much of this data has a specific audience: in our example, parents choosing a school for their child. Most parents will only face this situation a handful of times in their lives, and a large portion of the population aren't parents at all-- which means the data (though useful at those specific moments) is not as valuable to citizens and businesses as other types.

Planning data is generally used to make decisions about the future: either to change plans within government (by lobbying or voting) or to alter plans outside the public sector in response to government's intentions. An example of the second case is planning a journey by rail next week because road plans indicate engineering works that will make driving to the destination difficult. Another example is releasing a school calendar in electronic form, which means parents can have school holidays and sports days appearing in their own Outlook or Google Calendars.

Planning data can take some work to publish in a machine-readable form, because we are used to publishing much of it in documents with words, and without the addresses, geocodes or calendar formats, etc. that would give the data context and make it useful. This data can be fairly specific, but is of interest to a lot of the population who interact with public services or whose lives or businesses are affected by the decisions made in the public sector.

[1] 'There's data, and there's data.' Paul Clarke, HonestlyReal, 4 June 2010. http://paulclarke.com/honestlyreal/2010/06/theres-data-and-theres-data/

Infrastructural data is, to us, the most valuable data. It is the glue that ties together the disparate datasets published all across the public sector. What do rubbish collection routes and snow forecasts have in common? Though one comes from a council and the other from the Met Office, they both map back to specific points in geography-- which allows us to determine where snowfall may have an impact on bin collections.

Infrastructural data is also among the easiest for the public sector to publish, as the reputational risks are lower. Historical spending data, for example, has proven to be a source of headlines, sensationalist claims of government waste, and significant additional work on the part of Comms and Press offices. (This is notwithstanding a number of worthwhile and cost-saving uses for that data, but those tend to generate less media attention.) Infrastructural data is much less likely to prompt a scandal, and is therefore more palatable to talk about publishing.

Operational data is similarly useful because it affects the decisions we take, right now. If the train line is down, I must find another route. If my neighbourhood is being flooded, I need to know which direction is safe for evacuation -- which we may not be able to determine before the fact.

Ideally, this data should be freely available, but the costs can complicate the situation. Because operational or real-time data changes regularly (often by the second, minute, hour, or day), it can pile up quickly-- which can lead to an expensive hosting, distribution or publication costs. For organisations that must store this data (turning it into historical data, for example to determine the on-time performance of a train operating company last quarter), the costs may not be trivial.

We see business opportunities in making this data freely available for a short period of time, for example for 24 hours after it is generated. We expect that companies will come along, attach themselves to the feed of data and pay the storage costs themselves, charging users for the access to use the historical mass that they have accumulated. This value-added service is one possible way for the public sector to minimise the required investment in opening data while still making as much available as possible.

1.1.2 Licencing

We are creating a new market for this previously-unavailable data, which means that the creativity and expertise of the technology, research, science and media sectors are rising up to

interact with the new supply of data. We don't know what kinds of innovations they will find for various situations, what sorts of problems they will solve, or which datasets will prove useful ingredients.

As a result, we advocate as open licensing as possible, such that intellectual property restrictions will not provide a barrier to the potential innovations. We think Government should continue to encourage commercial (and non-commercial) re-use, and that the Open Government License provides a strong model for all public sector organisations releasing data.

As with the train operating companies holding data for live timetable adjustments, we have numerous situations where third parties generate data as part of delivering services on behalf of government or as commissioned by the public sector. The train operating companies, within the rights of their licences from the Office of Rail Regulation, retain the intellectual property rights for that data (and consequently charge a fee for its use). This creates a barrier for numerous entrepreneurs, developers and companies from producing products that could sell well, generating tax revenue and improving the travelling public's experience of the rail network.

As the Open Public Services agenda expands competition for the provision of public services, this problem is likely to pop up again and again. We would like to see the data generated by publicly funded activities freely available for re-use, which requires a change in procurement practices. Contracts with public bodies should default to ascribing the intellectual property rights of any generated data or information to the Crown to be subject to Crown Copyright and the Open Government Licence (or a parallel state of public ownership and open reuse licence, for the NHS and local government) as encouraged by the Office of Public Sector Information.[1] [2]

In addition to the work of the National Archives' Office of Public Sector Information, we have tried to build our view in echo of the Coalition agreement:

> *We will create a new 'right to data' so that government-held datasets can be requested and used by the public, and the published on a regular basis.*

---

1 See 'Crown Copyright: An overview for government departments.' Office of Public Sector Information, National Archives. http://www.nationalarchives.gov.uk/documents/information-management/crown-copyright-an-overview-for-government-departments.pdf

2 See 'Sample commissioning contract: Annex A to Copyright in works commissioned by the Crown.' Office of Public Sector Information, National Archives. This document exists to explicitly ascribe the intellectual property rights over a supplier's work to the Crown for publication and reuse. http://www.opsi.gov.uk/advice/crown-copyright/copyright-guidance/copyright-in-commissioned-works-annexa.doc

We will ensure that all data published by public bodies is published in an open and standardised format, so that it can be used easily and with minimal cost by third parties.

http://www.cabinetoffice.gov.uk/news/coalition-documents

1.2 Aspiration 2. The data should be as up-to-date, accessible, versatile, available and understandable as possible.

In a perfect world, all public data should be released immediately, in a variety of formats (CSV, XML, JSON, RDF), with thorough metadata explaining both:

- the dataset as a whole (the methodology by which it was created, the funding or sponsoring part of the public sector, the geographic areas covered, the time or date span, who to contact with queries, provenance or reliability of the data, links to previous versions or related datasets, etc.)
- the parts of the dataset (the meanings of each column heading, descriptions of what a row means, abbreviations and codes used, etc.)

Also, we would like the data and metadata to be provided either through a RESTful API, or posted on the Web-- but either way, it should be available at a URL that never changes. This persistence means that software and applications can continue to poll the same location for updated versions, resulting in less service disruptions for users relying on those applications. Similarly, previous years' data should not be deleted in favour of new editions; both may have value in different situations.

The data should be free of errors, typos, mistakes, and should standardise all vocabulary, making it easy to connect different datasets on different topics. As an example, a dataset with a column heading 'Council' might not show an obvious connection to one that includes a column for 'Local authority' (because the words don't match).

The data should also be easily findable (both from a search engine and from the relevant Department or public body's website).

1.2.1 LinkedGov

LinkedGov exists because the above requests are not generally feasible across the many data teams of the public sector. Instead of waiting for all of them to change, we are centrally organising crowds of entrepreneurs, developers, researchers and data experts to help with the reformatting, the cleaning and fixing of typos, the reconciliation (linking 'Council' to 'Local authority' and to 'LA' and to 'County'). We are also engaging with the publishers of data to

request, in a very simple quick way, enhanced metadata so that we can attach it to the datasets. And we are providing the resulting clean, linked data in a variety of formats with persistent URLs.

We, as a programme, exist to help with the work where the public sector doesn't have the capacity or tools to publish the data in the most re-usable ways. However, anywhere that the data teams are potentially growing, we would like to encourage them to move as closely as possible to those ideals (resulting in less work for us as a community and more immediately usable data).

1.2.2 Reliability

Not all public data is created equal: some datasets are backed by significant investment and statistical rigour (for example, National Statistics), while some are temporary or draft efforts (such as a spreadsheet from a Facilities team used to count desks on a floor, where they weren't bothered if the numbers were off by a few).

This range of reliability has already prompted some confusion in the data reuse community, and the potential risk for further confusion is significant. Rather than impose any kind of unrealistic threshold for publication ('This data must be 100% justifiable and true before it is released to the public') which could keep data from being published, we are instead working on communicating accurate expectations to the entrepreneur or developer.

We would like each dataset to be published with a raw Reliability score, indicating its level of rigour and credibility to the outside world. We hope to work with the Office of Public Sector Information at the National Archives, the ONS, UK Statistics Authority, the NAO and other relevant parties to achieve a standard scale (for example, 1-10). Factors currently under consideration for inclusion:

- known accuracy of the data
- how recent it is
- provenance: where the data came from, what journey it has taken
- methodology: how was it collected
- consistency of the data
- online availability of the data
- support for questions (either about the data itself, or the API/supply of it)

- standardisation with similar data from other sources in the public sector

This will help to create the re-use market, allowing businesses, developers and research to understand how to interpret the data they are using and how much importance to give it in the context of their own decisions.

We would be grateful for Cabinet Office's endorsement and encouragement in making this effort useful for and implementable within the public sector.

1.2.3 Policy initiatives

We have built our views around the Government's open data agendas. The Prime Minister summarised this in his letter to Cabinet Ministers on transparency and open data by saying:

> *In order to maximise the benefits of transparency, it is vital that data released by government are accurate, consistent and easily navigable. Over the next 12 months, we will take steps to improve the quality of data already being published, and ensure that it is updated on a regular basis.*
>
> http://www.number10.gov.uk/news/letter-to-cabinet-ministers-on-transparency-and-open-data/

And in his letter to government departments on plans to open up Government data:

> *Given the importance of this agenda, the Deputy Prime Minister and I would be grateful if departments would take immediate action to meet this timetable for data transparency, and to ensure that any data published is made available in an open format so that it can be re-used by third parties. From July 2010, government departments and agencies should ensure that any information published includes the underlying data in an open standardised format.*
>
> http://www.number10.gov.uk/news/letter-to-government-departments-on-opening-up-data/

1.3 Aspiration 3. Strong communication pathways between data publishers and data users should be established and maintained to help the public sector meet the needs of data users.

We are not only establishing new ways of working within the public sector, we are also enhancing and expanding a new market of data re-users in the UK. Both of these kinds of growth require continual adjustments, making sure that (in government) we are publishing as efficiently and effectively as possible, that we are listening and responding to those using our data, and encouraging the prime minister's agenda as best we can.

Similarly, we have found the greatest public data re-users to be in government, which means that they should also be helping to drive the publication process. Where efficiencies can be found, we should encourage them-- and accurate, relevant, usable data publication provides a significant opportunity.

Creating these communications pathways may require new skills for Press Offices, who generally responds to queries with set lines and explanations. They will need to understand the data their department, council or agency has published. They must be ready for questions along the lines of, "In cell C25 of the dataset for spending this quarter...", and they will have to provide contextual data in order to respond to the query. They may also establish links to the data teams (previously disconnected from Press Offices and Comms teams, in most public bodies) who can help inform them about the data being published.

These new roles for both the existing communicators and the data, information and technology teams will require resourcing, but will help data users enormously in learning how to think about what is on offer.

2. Questions for consultation

A selection of questions from the consultation, and our responses, are included here. Though we would have liked to provide comments on every question, the time and resources we had available have restricted us to these.

2.1 Section 1: Glossary of key terms

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

We are happy with the definitions given for all terms, but would like to expand the definition of Open Data (in relation to public services):

> *Data which can be freely used, re-used and redistributed by anyone.*
>
> *In relation to public services, Open Data means data available under the terms of the Open Government License.*
>
> *The presumption is that data **used in delivering** public services will be Open Data. It may be that some data held in relation to public services is made 'available', but is charged for, **where that data is generated by value-adding activities outside the scope of the organisation's public task.***
>
> (Our changes are marked by bold text.)

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

There are a number of criteria we would like to see applied across the public sector, when considering whether a dataset should be published:

- Is this data a part of the organisation's public task (and is therefore subject to the European Re-use of Public Sector Information Directive 2003)?
- Is the data free from any third party intellectual property restrictions or national security limitations?
- Does the data avoid identifying individuals in a way that violates the Data Protection Act?

- Is the cost not prohibitive? (This should be in line with the cost obligations set forth in the Freedom of Information Act and the Protection of Freedoms Bill.)

Unfortunately, resources may not always permit the publication of all data that fits those criteria. When we are forced to prioritise the publication of certain datasets before others, we recommend these additional tests:

- How much existing demand is there for the data? The more demand should result in a higher priority.
- How useful might this be to data re-users (who may not express demand because they may not have seen enough of it to know that they are interested)? Again, more potential should lead to a higher priority for publication.

3. If the costs to publish or release data are not judged to represent value for money, to what extent should the requester be required to pay for public services data, and under what circumstances?

The costs to release data should be minimal, if the data:

- does not need to be cleansed of personally identifiable data, third party data with a restrictive license or data which compromises national security before publication
- is already being held in an electronic format
- is already in one database.

Where infrastructure must be built (for example, APIs or report views for publication), these one-off costs should not normally result in an ongoing expenditure resulting in higher costs to data users.

We agree with the ethos of the EC Re-use of Public Sector Information Directive 2003 that if data is outside the public task of an organisation, it is fair for them to charge rates they deem appropriate to the market. If the data is within the public task, we would encourage the organisation to put in place the processes to work to lowering the marginal cost of producing that data to zero.

In the meantime, where it is not cost effective for an organisation to release a dataset, then it is reasonable to pass that cost onto the requester. (This should be in line with the cost obligations in the Freedom of Information Act and the Protection of Freedoms Bill.)

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 seems appropriate to use the definition from the European Re-Use of Public Sector Information Directive 2003:

> "1. 'Public sector body' means the State, regional or local authorities, bodies governed by public law and associations formed by one or several such authorities or one or several such bodies governed by public law.
>
> 2. 'Body governed by public law' means any body:
>
> (a) established for the specific purpose of meeting needs in the general interest, not having an industrial or commercial character; and [Does this rule out the PDC?]
>
> (b) having legal personality; and
>
> (c) financed, for the most part by the State, or regional or local authorities, or other bodies governed by public law, or subject to management supervision by those bodies; or having an administrative, managerial or supervisory board, more than half of whose members are appointed by the State, regional or local authorities or by other bodies governed by public law."

(EU Re-use of Public Sector Information Directive 2003, Article 2: Definitions)

2.2 Section 8: Policy challenge questions

An enhanced right to data:

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

The PM and Government policy have already made it clear that there should be a strong presumption of publication. It is the implementation that is lacking, or possibly the understanding

that the default presumption should be to publish. Because of this, what is needed is a communications program for the bits of government that have not received the message yet, and possibly central point where people can report difficulties obtaining data. In order for it to be effective, the people resourcing that central point would need to have appropriate skills and influence to be able to help solve the problems raised.

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?

This body should be outside of political control, accountable to the people, and endowed with sufficient power to ensure compliance with strategic goals and agendas set. The body would need some technical skill, because it would have to be able to look at the data and check that the information that was claimed to be in the data was actually there and in a usable form.

We have found that compliance with instructions to publish can end up with a conflict: where an organisation may decide, strategically, to release a dataset, the publishing team may not have the resources or skills to make it usable. In practical terms, this means that the feed from an API may be missing half the data, for example-- resulting in developers struggling with incomplete information and publishing team nonetheless reporting to senior management that they have published an API. The body described in this section must have sufficient technical skills to understand what data is there, and enough political power to encourage change when it may need to be catalysed at the highest levels.

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

No. We believes that anonymised or “de-identified” data is risky to the success of government’s vision for Open Data for two reasons:

a) The risk of de-anonymisation is significant, potentially putting citizens in danger or damaging their reputations. We must take care, also, to heed Article 8 of the Human Rights Act 2008 on the right to respect for private and family life, which states:

> *Everyone has the right to respect for his private and family life, his home and his correspondence.*

There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others.

http://www.legislation.gov.uk/ukpga/1998/42/schedule/1/part/I/chapter/7

Government is fragmented, and cross-referencing data from different organisations can make it easy to work out who is whom. For example, if anonymised health data gives sensitive information about a 70-year-old man in Leicester with psychotic episodes, headaches and prior exposure to a toxin used in warfare, we could possibly identify the man by:

- working out his years of military service from the era the toxin was used, and cross-referencing any online lists of military veterans (perhaps a local club)
- searching social services data to find a nearby council looking after people with psychosis, and seeing how that overlaps with the catchment area of the patient's hospital or GP

At a minimum, this would restrict the list of possible patients down to a manageable number for searching one-by-one, if someone were determined. At worst, this could fully identify the patient and betray the their confidentiality.

It is a high demand (costly, etc.) to ask public sector bodies to scan the landscape of data available from other organisations and the private sector to adequately anticipate the dereferencing possibilities.

b) Because personal data is hugely important to individuals, and because it cannot be clawed back once released, the political risks in publishing anonymised data high. Any public embarrassment could not only cause the relevant public sector body significant trouble, it could also reflect badly on the Open Data and Transparency agendas and possibly impede the publication of other open data.

For both these reasons, the release of anonymised data, as a topic, must be approached with extreme care.

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?

Without knowing the extent or nature of the data held, nor how it is currently being stored and managed, it is impossible to determine the resources that would be required to satisfy the enhanced right to data.

At a minimum, each body will need a route via which they can publish the data. They will also need to have the skills and resources to do so in an appropriate format. Often, this will simple mean using an option that's already available in the software used to manage the data, for example, data from an Excel spreadsheet could be exported as a comma separated values (CSV) file.

Government itself is one of the heaviest users of government data, yet often it is not as easy to access even from the inside as one might expect. Much duplication of time, money and effort would be saved simply by making the data more easily available. This freeing up of resources may help to offset some of the publication costs.

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

We would like to see the publication of open data as a requirement (ideally, as close to what we have laid out in the 1.1.2 Licencing section as possible) included, by default, in every procurement contract. Most new ICT systems, with a bit of planning, can easily provide appropriate data feeds to outside users -- but adding in this requirement at a later stage may prove more costly and potentially disruptive. Similarly, an assumption of publication helps us to get as much data out as possible; if we wait for someone to request that data, we have to similarly allow for them to discover that it exists. This could take a while, potentially hampering innovations that could be built from that data in the meantime.

Similarly, as set out in section 1.1.2 on Licencing above, where a supplier is providing a service for government, the procurement process should default to (and make explicit) the intellectual property rights for all generated data being published under an open, non restrictive (commercial or non-commercial) licence for reuse.

Setting Open Data standards

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

Ideally, standards should be set by a consensus view of everyone using a specific kind of data. Where consensus is not possible or feasible to achieve (as is often the case with data standards; too many stakeholders are inaccessible in most cases), data should be as interoperable as possible, making it easy for individual users to impose their own standards.

To encourage compliance where standards have been agreed, we encourage government to set the standards burden very low. Raw data is the most interoperable kind of data, and reformatting may not be an insurmountable burden for data users. The less additional work we ask of our (already overburdened) data teams, the greater chance we have of compliance.

If standards are adopted by the public sector, they should be openly available and free of charge for the public sector to use in publication and for users of the data to implement to make sense of it. Ideally, they should be openly and fairly developed as well, like those from a standards body such as the World Wide Web Consortium (the W3C).

If a standard is adopted by the public sector, relevant regulators should have the power to enforce compliance with it.

Meaningful Open Data

5. Should the data that government releases always be of high quality?

No. The burden on government is too high to improve every dataset (and often unnecessary; for many situations, an approximate value is as useful as a specific, perfect one). The important thing is to communicate to businesses, developers, or anyone using the data the level of quality, so that they can set their expectations accordingly. See our above discussion on Reliability, section 1.1.2.

Government sets the example

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

Overall, our goal is to publish as much data as possible. Where we must prioritise, we would like to see the decisions based on:

- Demand. If there is any demand for a given dataset, this should be prioritised. A comprehensive and easily accessible asset register would help to facilitate demand.
- Disambiguation. Datasets that contain indices or schemas that are referenced by or adhered to by other datasets around and beyond government. This includes infrastructural or core reference datasets, such as Ordnance Survey's Code-Point dataset which connects postcodes to geocodes.
- Currency. Live data should be given a high priority as its usefulness may degrade swiftly as it ages. As also should any data with schedules and timing of things that the general population could be interested in.
- Cost. Any dataset with a low or negligible publication cost (or resource effort) should be published as soon as possible. This may be before those meeting the criteria above.