emapsite™

Response to Making Open Data Real: a Public Consultation

From: James Cutler

for and on behalf of emapsite.com Limited

Background

emapsite.com Limited was founded in 2000 and will turnover £3m in FY2011-12, generating a taxable profit and employing 16 full time staff. As an online location content platform emapsite is built and continues to grow thanks to the very best of UK talent across many different disciplines. Beyond doubt and with plentiful experience and evidence emapsite represents what it is possible to achieve if an entrepreneur and their team identifies clear market segments and develops sustainable business models that service those markets. emapsite has invested in, and continues to invest in, a diverse network of relationships with more than 30 location content suppliers and in the infrastructure to support, manage and maintain a platform that provides their data to our users.

emapsite's interest in this debate stems among other things from the fact that one of our suppliers is Ordnance Survey with whom emapsite has been a Premier Partner for over a decade.

Amongst emapsite users are system integrators, public sector entities and over 5000 commercial end users who take content, including open data and Ordnance Survey data, from emapsite via a range of services and interfaces, for use in their own applications.

As is evident from this response emapsite fundamentally supports an open data agenda which demands that government release, free for re-use and re-distribution, all data generated by government for government (sometimes termed "exhaust" data or data about the business of government), subject to a number of tests. emapsite agrees that such release will broadly but to varying degrees assist in developing the transparency, accountability and growth agendas and that this will take time and require government resources. emapsite is concerned that some participants in the open data agenda wilfully conflate such open data with data that does not satisfy the proposed tests, notably but by no means exclusively data currently traded by Trading Funds and other agencies of government. In particular, given the extensive release in 2010 of OS OpenData™ (including postcodes and much else besides), emapsite believe that this diverts all participants from the challenges and opportunities that exhaust data provide.

Response to Making Open Data Real: a Public Consultation

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

   The definitions perpetuate the conflation that has dogged this debate for some time. In particular and despite effort to the contrary, data, dataset and information are used interchangeably within the glossary and from then on. Therefore, I would urge that due reflection is given to what terms should be used and how and that these are then persisted with.

   However, the key definition, that of Open Data, only needs a tweak to reflect the definition of data set to provide a basis for progress and clarity.

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

   The simpler the tests are the better. The basic ones are:

   - Would the data be released under a valid request from any of the methods available within the regulatory framework, namely the Freedom of Information Act, the Data Protection Act and the Environmental Information Regulations (and others)
   - Is the data a by-product of public sector activity or otherwise collected by the public sector for monitoring and evaluation of the performance of that public sector activity
   - Can the data be delivered under the Open Government Licence (OGL)
   - Are other data sets available, be they open data, web scraped or commercially sourced, that facilitate deanonymisation of the data set to be released

   If the answer to the first two is positive then in principle the data set concerned should be open data and the OGL applies. If the first two answers are positive but the OGL cannot be applied (typically for reasons either of sensitivity (personal, national security or damaging to public finances) or of shared/commercial/third party IP in the content and the fear that such content can be reverse engineered), then there may be good reasons why the dataset should not be released in the proposed form. By extension if the answer to the fourth test is considered to be (and can be demonstrated to be) in the affirmative then the data set should not be released.

   The third test is the most sensitive in that some (many even) data sets will contain elements of third party data each subject to different licensing. The most notorious of these is Royal Mail and Ordnance Survey IP within postal and other addresses.

However, with the continued expansion of the role of the private sector in public service delivery it seems inevitable that this will become a greater challenge as competitive issues and instincts will be used to filter what can be released. In this regard it is essential that those delivering pubic services should be covered by the regulatory framework in the same way that pubic sector agencies are themselves covered. This is not currently the case and it seems likely that fragmentation of public service delivery will make data set aggregation ever more challenging unless standard metrics for reporting and data set release firm part of the contractual arrangements with the private sector. Unfortunately it is easy to foresee great resistance on commercial grounds to such developments.

The more technically difficult test is the final one; are PSIHs or should they be qualified or resourced to make (or have made) an assessment of such feasibility? It seems increasingly likely that this will become a necessity if Government is to fulfil its obligation to keep personal data private.

Is there any evidence to suggest that requests are made with the aim of integrating the resulting data set with third party data to assist in activities that include deanonymisation? There is probably little hard evidence to confirm or deny such a proposition but it would be naïve to suggest that it is not plausible.

Thus, if the answer to the fourth test is to err on the side of caution then some data sets may well not be released under such requests. This would help limit release not only of personal data but also of other data that can either be deanonymised[1] (to the detriment of the person, people or entity otherwise protected) or otherwise exploited by third parties, be it for frivolous, commercial or social gain.

The current mechanisms could be adapted to place the burden of proof on the PSIH both in terms of the reasons for the ineligibility of the data for OGL (straightforward) and the likely risks for deanonymisation (more challenging and a risk in that the PSIH could use this to stonewall every request). This would require the PSIH to better understand and document their own data assets. In addition the PSIH could be required to provide evidence of the deanonymisation risk and an example from the requested data set that has been adequately de-risked.

[1] There is plentiful literature on deanonymisation; from recent ICO presentations www.ico.gov.uk/~/media/documents/anonymisation.../ohara_slideshow.pdf to more fundamental technical assessments: http://randomwalker.info/social-networks/index.html, http://33bits.org/2011/03/09/link-prediction-by-de-anonymization-how-we-won-the-kaggle-social-network-challenge/, http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf,

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?**

As I understand it, the regulatory framework (FoIA, DPA, EIR) provides scope for the public sector information holder (PSIH) to recover such costs should the need arise. This still leaves open the question of "value for money" and opens up the possibility for abuse. Even inserting a scale of charges does not defeat this potential barrier. In addition the request can be exploitative in as much as the request may in effect be for "information" i.e. data in a value added form that would make sense to or ease interpretation for the requester. It is not the duty of the public sector to perform such value adding services beyond that which might be found in published reports. So, the scope for conflict will remain unless some additional tests can be found.

I suggest that the "test" be some form of rapid technical assessment of the nature of the data requested.

> By way of example, if it found that the data is in a single file and can be easily "exported" then the charge might be nominal or "free".
>
> In contrast the data requested could be for example spread across numerous databases and/or in different offices or departments. Then the request is essentially one for data aggregation, assembly, cleaning, normalizing or other form of data processing that might in a commercial context be considered value add and does not form part of the day-to-day requirements of the PSIH, then the PSIH should seek to deny the request or charge full cost recovery for each element and fulfil the request as a series of "downloads" or "exports" that place the effort of aggregation, linking and interpretation entirely with the requester.
>
> If the data is deemed outside of OGL then the third party licence holders need to be documented.
>
> If the data is deemed sensitive to deanonymisation then a sample "safe" data set should be released.

However, the PSIH must not be obstructive and should be expected to provide accompanying metadata for each data element. In due course it is expected that PSIHs will proceed down the 5-step 'linkeddata' path, albeit at some considerable (if not yet recognised or considered) cost[2] but with the objective of reducing the long term cost of data management and inherently increasing accessibility.

[2] EDINA's experience as evidenced at AGI geocommunity 2011 indicates just how hard this path is: (http://assgeoinf.squarespace.com/storage/AGI%20Conference%20Guide%202011.pdf, paper yet to be published)

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?**

   Any data that satisfies the "Open Data" tests above should be in scope in principle. However, as noted in answer to Q2 above, government in general finds itself in an inherent conflict given the level of outsourcing, PPP and PFI involved in execution of public sector policy objectives.

   While such relationships are competed for on a level playing field, they will be subject to all manner of negotiations in the final terms and the subsequent execution of the contract. Similar tensions exist in housing associations, educational and hospital trusts and all manner of areas where "subsidy" in its myriad forms plays a part in the operations of the public sector framework (transport, defence, emergency services, justice, private sector R&D and so on).

   This is a far more complex area than might initially appear. The public sector "compact" provides huge flows of money across the economy and should not be used as a bat with which to enforce the debate.

   Arguably what is genuinely required is the establishment of metrics and transparency requirements for each sector or PSIH or combination thereof and which apply to PSIH or executing entity at all times, forming if necessary statutory and contractual reporting protocols.

   If this were initiated, any requests that would satisfy the tests in Q2 above were they to be applied to PSIHs (including those executing public sector policy on behalf of government be they commercial, voluntary, charity or third sector) would merit the immediate release of those metrics (were they not already published), together with appropriate contractual data and the underlying "raw" data.

   While many sectors already have the metrics, the protocols for the underlying contractual and "raw" data are much less well-formed (or resistant to formation through either wilfulness or (alleged or actual) complexity). Nevertheless fragmentation of commissioning and execution makes meaningful data aggregation more challenging in some sectors.

   The bottom line is Government and the regulators must apply far more stringent data publishing requirements on PSIHs and those that do their bidding, publish what they should be (which would assist identification of data that should (and should not) be open) and hold a mighty enough armoury of enforcement to ensure compliance.

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

I believe that the ICO already has adequate powers and that the mandating of data release via data.gov.uk is seeping across the PSIH body, albeit at varying speeds. Could PSIHs act more quickly? Probably. Do they have other priorities and limited budgets? Yes, and it is hard to argue the case for open data when efficiencies need to be made.

I have mooted[3] that a Public Data Corporation is exactly the kind of entity that could create best practice and other guidance with case studies and standards references that would help accelerate this process with a light, progressive touch.

The Public Data Transparency Board (PDTB) is laden with good intentions around the formalization of the semantic web as the approach with which to achieve these goals. There are of course other ways to achieve the same ends as the Open Geospatial Consortium (OGC) has realised in accommodating quasi-proprietary submissions made by commercial entities such as Google and ESRI. The PDTB needs to be alert to what is still an evolving domain and encourage diversity of solution adoption as long as alternative approaches do not limit inter-operability.

On a different subject, there is at least one reference model for sectoral or vertical market data aggregation and publishing from which lessons might be learnt. The NLPG, for good or ill, was set up as a hub to create a standardized data set from 350+ sources. It cost a fraction of the Royal Mail's overhead for the postcode address file (PAF) to set up and maintain. It is not inconceivable that a similar conceptual approach (technical solutions have evolved somewhat) in other sectors or activities might achieve similar advantages and efficiencies of scale.

In the meantime myriad competing SMEs, vanity publishing projects, students, bedroom coders and others are busy replicating efforts to assemble sector or activity specific data sets. Which brings the focus back to efforts to establish minimum expectations and standards for capture and publishing of metrics, contractual and "raw" data e.g. for transport timetables, for crime, for emergency services, prisons and so on.

Meanwhile, it seems appropriate to mention that the world's leading statistical body, the ONS barely ever gets a mention and is well placed to advise in this process.

[3] In a previous submission to the PDC consultation in March 2011 (copy available if need be)

Some general observations about the consultation document:

Section 3, para 3.2: "better" data does not mean "less" data; indeed we live in a data tsunami and the volume of data being collected and that can be collected even within the PSIH environment continues to increase. The potential for "big data" to provide previously unavailable insight through aggregation and analytics and to actually, over time and with resources, lead not only to an increase in data quality but also to the more informed identification and analysis of outliers is one of the hopes for overall performance and efficiency gains.

Section 4, para 4.3: very selective quoting of figures; a more recent cross-Europe analysis suggests a far more prosaic figure and even has qualms about that: (http://www.epsiplatform.eu/news/news/review_of_recent_psi_re_use_studies_published).

Section 6, para 6.9: this paragraph seems to suggest that it is acceptable for PSIHs to embark on value adding services; this is palpably not the case, be it for data cleaning for FoIA requests (see above) or as per the restrictions on Ordnance Survey in the PSMA. The private sector is the proven environment for innovation and the creation of agile, flexible and cost effective services; what they need is data, not a public sector that competes. This is not to say that PSIHs shouldn't (themselves or through contractual arrangements) publish "derived" or "value added" "reports" on their activities to ease interpretation and communication with the majority of stakeholders. Rather it is to ensure that they do not do so on a competitive or market making basis.

Section 6, para 6.11: this paragraph suggests that the authors consider that there is "unnecessary" data collection. Not requested does not equate to not necessary; indeed as the data capture environment becomes ever more ubiquitous more and more data will be collected in any case and as it will be attributed with metadata and in digital form from the outset, individual elements will be discoverable and accessible. "The smartest thing that will be done with your data will be done by somebody else" becomes ever more true as the volumes of data expand and the tools to analyse that data become more established; do not throw out the baby with the bathwater.

Section 6, para 6.12: unless I have missed it, the report referred to has not yet been released.

Section 7, para 7.7: citing volunteer data portals, however useful and admirable, is to completely ignore the reality that they are an unsustainable temporary public good.

Section 7, para 7.8: this is overstating the case; there is a dearth of evidence surrounding the area of economic of public sector information. It would be wise for the consultation to acknowledge this plain fact and to reach out to the community for more evidence. While there is little question that efficiency gains and social gains

can be found by learning from PSI, it is at best unclear as to what financial and wider economy gains there can be through the exploitation of that data at a time when growth is essential.

Creating services for the citizen from open data to provide more accessible insights is already a highly commoditised and competitive market place. The remaining B2B marketplace is central to the creation of a thriving economy and is distinguished by the value it places on provenance, currency, quality and other factors, which accounts both for the success of the big data companies and the fact that those attributes cost. That is to say the real wealth creators while always content to take something for nothing also recognise that there is a cost of sale and that the payment of that cost imbues the supplier, amongst other things, with the ability to persist to supply that data at the required quality.

The other significant market for PSI is the public sector itself. Taking PSI and creating services to be sold back to the public sector could yield significant savings if for example projections in the healthcare sector are delivered. In healthcare, some of these savings will come about as a result of data that is intended to be 'open' but much of it is also predicated on somehow harnessing personal data, something outside this consultation (as should thus be the financial and economic implications).

The benefit to UK plc of a healthy, established data environment cannot be underestimated; these could be rapidly undermined if that environment is damaged through the placing of unnecessary restraints on "raw" data capture or placing shareholder dividend ahead of the national interest. Therefore, data capture, storage, maintenance and sharing demands continued nurturing and protection across PSIHs from agencies whose data is 'naturally' open to Trading Funds.

Policy Challenge Questions

A An enhanced right to date

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

A definitive link between the presumption and the definition needs to be established and then embedded in training (induction, in-service, as a metric in annual reviews and appraisals etc), in manuals. That presumption should be as per answer to Q2 above, that business of government data is open if it passes those tests.

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?**

As per the point in my answer to Q4 above, the key is to have a range of regulatory powers that can be used and are seen to be used. Those that wield the powers, be it PDC, ICO, OPSI, OFT, TNA or other need to be knowledge and practice leaders in this area, providing guidance, case studies, exemplars, support, templates and the like as well as an 'accessible' sounding board (perhaps incorporating PDTB) long before formal enforcement measures are required.

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

Very unlikely. Apart from damaging "lost" and "stolen" events this issue is perhaps more central to the open data debate than many yet think or are willing to admit. Kieron O'Hara's report and the plentiful technical research in this area (referenced above), particularly around deanonymisation, suggests that there are serious challenges. As noted above (Q1) it seems increasingly necessary that Government needs to understand the risks to privacy that deanonymisation represents and introduces protocols to the appropriate regulations to increase protection for personal data in the face of a commitment to otherwise open data.

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?**

In the short term at least there could well be an increase in costs associated with processing FoIA and other requests. In the medium term resourcing Government with skills both to deliver on the promise of the 5-star linkeddata approach and to protect personal data is going to require considerable resource investment, primarily in

personnel such as data scientists. If any economic benefit is to come from open data then this investment is a pre-requisite and will at any level be proportionate to the overall aims of an open data policy.

As yet there is no visibility (transparency) of the cost to government of the pursuit of the open data agenda thus far. While perhaps "small" in the scheme of things there is plentiful anecdotal evidence of the time and resources being spent by or within the Civil Service in pursuit of this agenda (illustrating once again that there really is no such thing as a free lunch) which arguably could be better or differently spent in the current climate.

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

Simple, transparent steps:

- define the acceptable standards, protocols and approaches;
- mandate them in terms of service, contracts and elsewhere with a PSIH specific series of milestones;
- disqualify tenderers who refuse to comply;
- discipline personnel who fail;
- publish milestones and performance;
- penalise PSIHs for performance against the agreed milestones for adoption.

Comments

It is of concern that "rights" are increasingly seen as an entitlement and a moral absolute. This becomes particularly unsettling in a national or global context when it is unarguable that issues such as poverty, preventable disease and education to name but three are not themselves adequately resourced nationally or globally and that resources could be diverted from many sources to contribute to genuine improvement. Therefore, to assert a "right" to data, whilst understandable in the narrow context of open data, transparency and so on is to ignore the bigger picture. With civil servants themselves, via social networks, suggesting that huge amounts of resource are already being committed to this effort and with researchers such as those at EDINA suggesting that the road to linkeddata is paved with very significant potholes and other dangers, it does beg the question as to whether any of this should be a real priority given the current economic conditions.

Nevertheless, in answer to the broad policy ambitions of clause 8.6 then the answer to the question as to whether change in the listed areas would assist in creating a healthy, flourishing and sustainable public sector data environment, is a qualified yes. The devil as ever is in the detail as noted above with respect for example to deanonymisation of personal data or to the long term benefit to UK plc of a high quality data resource across the public sector spectrum from central government department to executive agency or non-departmental public body to Trading Fund.

B Setting Open Data Standards

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

As above, simple, published steps:

- Agree standards, particularly for inter-operability (noting that semantic web approaches as promulgated by Sir TBL in the 5 –star rating approach are not the only solution)
- Publish them
- Agree milestones with PSIHs
- Create accessible resource base (templates, guidance, use cases, best practice, cost benefit analyses etc)
- Publish metrics regarding success in achieving milestones
- Penalise persistent failure to achieve milestones

Although many open data advocates would consider data released as PDF to be "closed" it is important to consider the citizen in this debate and the citizen is not versed at all in the arcane language of open data. And while it is entirely plausible that in the longer term citizens will consume all their services from intermediaries encumbered by advertising, subscription or transactional models in the short-medium term the release of data in PDF report form, be it by health or educational trust or government agency would by 99% of citizens be viewed to be considerable "progress". So we need to be careful about the balance of "opinion" in this debate. As Dr Foster, SpikesCavell and others have demonstrated over the last decade it is perfectly possible to develop solutions to the diversity and awkwardness of public data sets in pursuit of an identified and sustainable market without machine readable data (though of course it is desirable). Today's open data advocates do not all seem to apply the same rigour, research and market development energies and appeal to "free" data as a substitute.

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

User experience is not a consideration in moving along and meeting the 5-star rating approach, nor should it be.

The aim of this approach is for PSIHs to adopt a data publishing model for approved open data that enables those who wish to, to engage and interact with that data at their discretion. This inevitably means that the whole open data initiative will only satisfy the immediate demands a small cadre of technocratic campaigners as the skills required to engage and interact are relatively sophisticated. The failure of information asset registers in general land the lack of adoption of the 'unlocking service' coupled to the

absence of truly valuable (from social and economic gain perspectives) applications built on data from data.gov.uk or Ordnance Survey OpenData™ suggests that this is as yet a game played by the few.

The expectation is that this community will create new products and services for citizens, business and government that will contribute to the economy through jobs, taxes and so on. Given that the evidence for this expectation is negligible (in that most "applications" already using open data are vanity projects or subsidised by grant funding with little revenue from their audience or advertisers to justify the word "business" let alone profit, jobs and taxes) then it could be argued that by engaging with the intermediaries rather than the end users, citizens or stakeholders that Government has been lured by a small, vocal and high profile cabal down a route that fails to serve at least part of Government's role, that of communication.

As noted above there will be an inevitable requirement for data scientists and similar skills within government and it is anticipated that when government understands the potential of "their" data to improve their own services and performance that they will invest further in such skills and perhaps also in interfaces that demystify the interaction with the underlying data. Many open data advocates do little to encourage such within government behaviour extolling them to publish RDF and SPARQL end-points to the technorati rather than develop toolsets for the citizen.

If Government does the latter then user experience will become important but until government crosses that intellectual threshold open data will be only open to those (few) with the skills to find it, access it, mash it up with other data, drill down into it and republish it ass information. This does not seem to represent a rebalancing of the knowledge economy but rather the establishment of another knowledge silo.

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

There are already Government entities that have accreditation systems, from Trading Funds such as Ordnance Survey and the Met Office to Executive Agencies such as the Environment Agency to research organisations such as the British Geological Survey or even CLG's Planning Portal.

It is emapsite's belief that a number of cross-cutting goals could be achieved from these organisations within government:

- That such agencies can establish and operate a channel model in tradeable data
- That the market responds well to Ramsey pricing for such tradable data
- That tradeable data charges has fallen in real terms over the last decade in the absence of RPI linked (or indeed generally any) charge increases, by as much as 40%

- That such agencies are not cost-effective in building and operating a direct sales model except perhaps to government itself
- That efficiency savings across this spectrum of organisations could be achieved very rapidly through abolition of their direct sales and marketing teams, winding down of services that compete with commercial providers
- That such action would strengthen these agencies through ensuring focus on data quality and content, something that markets have been critical of
- That such action would drive down or at the very least stabilise long term charging structures for tradeable data providing intermediaries and users with the confidence needed to develop and adopt solutions and services that embed that data

C 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.**

See above – mandate, publish, support, measure and, ultimately, penalise.

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 is properly protected and that privacy issues are met?**

Public sector "boards" are few and far between outside those who have a trading activity. However, each PSIH should appoint an SRO or equivalent and all PSIH staff need to have within their terms and conditions a requirement to meet the above mandate as directed by the SRO. The SRO should be responsible for agreeing and then meeting milestones and should ultimately be in some civil service way be accountable for both success and failure in delivering the open data vision for that PSIH.

Protecting personal data is an increasingly technical challenge and the responsibility to ensure that that technical/analytical responsibility is fulfilled (through internal resourcing or external appointment) should fall under the remit of the SRO.

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

As above, yes, although the stick and carrot of agreeing milestones and having the success in achieving them measurable and measured and published is likely to be good enough except for the most intransigent PSIH.

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

The question suggests that a whole series of quangos be established harnessing sectoral expert knowledge. This seems an unnecessary and costly overhead for the open data agenda to have to accommodate. If the responsibilities and requirements are mandated and the milestones agreed on a PSIH by PSIH basis then there is a framework for open data release that is auditable, transparent and ultimately subject to penalty. So the direct answer to the question is no; however, establishing a PSIH specific programme for implementation very much equates to a harnessing of PSIH data knowledge in pursuit of the wider goal.

D Meaningful Open Data

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

From the above it is evident that this commentator believes that PSIHs would benefit significantly in terms of the direction and speed of travel towards an improved open data environment if they were provided with as many tools and as much support as is possible in determining and achieving their milestones. So frameworks with common language, consistent terms, examples, generic guidance and best practice are a worthy ambition. And data.gov.uk provides a good a portal as any for data publishing. The quest for Government is to establish an appropriate entity that can deliver these frameworks; the Public Data Corporation offers one option while equipping an existing entity such as ICO or TNA with the appropriate resources and skills might offer another and demonstrate the UK government's commitment to the open data agenda.

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

Simple answer, it shouldn't – the data should all be in an inventory whether it is open or not. Government, and even the PSIHs themselves, don't know the answer (as to what "should" be a priority – "the best thing that will be done with your data will be done by someone else"). An individual PSIH might be able to suggest a programme of release of its data and it may even do so in an order commensurate with value, user interest and so on but this is one area where the mandate must be around inclusion in the inventory. Said inclusion likely merits the documentation of that data set in terms of the tests suggested above (and others) and thus of its visibility or viability for inclusion/exposure/release via data.gov.uk.

Once again, guidance might be of assistance in informing and expediting a PSIH's open data programme owing to the diversity of "value" in end users and intermediaries eyes:

- Currency
- Accuracy
- Timeliness
- Granularity
- Format
- Interoperability
- Purpose
- Price
- Licence terms
- Coverage
- Completeness
- Provenance
- Relevance
- Metadata – level of detail, quality

- Tangible quality
- Linkeddata level and ability or ease of linking with other data
- Frequency of update
- Compliance (i.e. is it mandatory, does it comply with relevant legislative framework)
- Ongoing availability

It really is not possible for a PSIH to second guess data utility but rather to focus on complying with the open data agenda in the most complete way possible (in the long run).

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

Every public sector activity merits open data of some form, that's what transparency and accountability mean. Anything less would be hypocritical. Of course there are privacy and national security concerns that need to be accommodated but everything else should be considered potential open data. Many data sets' utility comes from their timely release and from subsequent continued availability for monitoring and evaluation, comparative studies and so on. It is very important that data set release is not seen as a one-ff activity but rather is embedded as an on-going part of PSIH data capture, storage, management and publishing activities. These are almost entirely automatable. See also answer to Q2 on page 1 of this submission.

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

None so don't try to identify them. Collection of all data should continue and preferably expand based on sensor webs, ambient capture technology and so on; data storage is very cheap, data mining and analytics tools are improving all the time, machine learning and other advanced computational developments permit intelligent interrogation of even the largest data sets offering untold benefits in terms of data cleansing, processing and so on and in terms of for example identifying and understanding outliers in data sets, often the key to breakthroughs in understanding.

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?

While it is desirable from both producer and consumer/user point of view that data collected and then published is perfect (for the purpose for which it is captured), to err is human so there are many potential sources of error from simple transcription errors to methodological issues that impact "quality" (relevance, accuracy, timeliness, accessibility, interpretability, coherence, precision etc – see Q2 in this section) of data and of metadata. Therefore, one can assume that "raw" data will contain errors of some kind, some of which can be identified during the journey from point of capture to point of use/publication/release.

That journey is the "polishing" to which the question refers and it is really the degree of polishing that is at issue here. This ranges from simple validation (such as double entry) and quality assurance procedures (for example for ISO9000 or other accredited or regulated compliance) to further data processing, cleaning, normalising, information derivation, analytics, anonymisation, re-districting, degrading and republishing to name but a few. The "many eyes" paradigm suggests that the more people involved in various aspects of 'polishing' the better the data will, over time, become, based on feedback to/within PSIH. On that basis publish all, publish early (but of course publish carefully re privacy/personal data). Once again, it is important to labour the point that data that will be 'open' is generally data about the business of government and should in theory at least serve those ends and not, through any 'polishing', those of third parties.

Comment

This section of the consultation document (especially 8.15) asks whether four specific approaches will help the user identify what it is that is available and whether it will be of any interest or use to them. The response to Q1 above indicates how these should be adopted. However, it is important to highlight a number of conflicts raised in this section, notably about priority, value and necessity. As per response to Q2 in B above, open data is data pertaining to the business of government and may or may not be of utility, value, importance or otherwise to any external constituency. Sure, that constituency might have an idea as to what it would like and why and what they could do with it and who for and how much they might even pay for the information product to be created, but it is not the role of the PSIH to analyse or second guess what that might be. As soon as that were to happen the PSIH becomes embroiled in a competitive arena with pressure on resources, lobbyists, increased FoIA requests and so forth. Value as they say will out and there is absolutely no role for PSIHs in valuing, filtering, curating, prioritising or in any way processing their data let alone market making.

E 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?**

Government should be seen as "just" another end user of open data and should be able to avail of it through the outlets adopted by PSIHs and Government more widely, notably data.gov.uk but likely others including PSIH-specific machine-readable feeds and APIs.

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

This is the "value" question again (see Q2 in D above) only using different words, implying that data should be released as open data if it will provide national social and economic benefit. There is plenty of evidence to suggest that "the best thing that will be done with your data will be done by somebody else" and that Government has no role to play in second guessing what data sets should be prioritised for release as open data. If the data passes the tests as open it should be released, it's that simple!

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

See above and elsewhere, Government has no role in prioritising.

Comment

Clause 8.17 makes some sensible suggestions (routine publishing of evidence and databases and underlying data for example) regarding how government might improve transparency in policy making which should become matters of principle for every government department, NDPB, executive agency or other arms length entity including local authorities, housing associations, educational and health trusts and so on.

F 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?

Government and innovation are traditionally uncomfortable bed-fellows in part because unlike 'research', innovation implies or invites the notion of direct commercial involvement carrying with it inference of seed/grant funding that sustain start-ups and other in the absence of a business model, market or other business fundamental. In this context government's role is far more to provide the environment in which particular kinds of typically high value add (manufacturing, biopharma, defence being examples) innovation can flourish than to seek to stimulate specific activities. Therefore recent initiatives such as the Innovation Launch Pad, the mooted product surgeries for SMEs and appointment of the Crown Commercial Representative for SMEs all fertilise that environment. Of course, such things represent an investment particularly at a time of public sector purse tightening so they need to be balanced against more nebulous demand such as those cited in clause 8.22.

Open data is itself a many splendoured thing with user communities ranging from individual citizens to whole regions or vertical markets. Public service providers should focus on their core remits and worry less about "promoting the use of data" as long as they have fulfilled their mandate to publish the open data that they generate or are otherwise responsible for. Anything else is to load up the public service provider with a value adding role other than that for which they collect that data in the first place and to do this would not only add costs but also establish new tensions around where value lies and whose open data priorities should be addressed first, or at all. The level playing field that is satisfied by releasing the data covered by the tests proposed will mitigate the extent to which this can happen.

Comment

Hack days and the like are beloved of the developer community but rarely result in little more than vanity projects with little or no take up, commercialisation or sustainable social and economic gain. The only real mechanism by which a wider audience can be reached and subsequent wider use of open data engendered is by making 'access' easier and while the ambitions of linkeddata may to some provide one solution, more easily understood approaches are already available. There is likely a role for government and for value adding intermediaries in evolving solutions including data portals with application programming interfaces that make it easy to use open data. As long as the PSIHs do their job around releasing the data then the value adding community, innovators and end users and well as developers will be well served. The commercial sector has not to my mind been fully engaged with or by these developments, allowing the developer community to drive the open data agenda. Government needs to recognise that the commercial sector invests substantially in activities to realise value around their core offerings; where that includes developing services, be it data cleansing, statistical analysis, reports, real time reporting or advanced analytics, innovation is to be found and economic gain, in profits, taxes, jobs etc are to be found.

Annex 1

Clause A1.56 This commentator would certainly like to be party to the work stream of Growth Review that is/will focus on and bring more depth to the economic benefits of Open Data y assessing the size of the opportunities.