SAM SMITH RESPONSE (VIA E-MAIL)

Dear Cabinet Office,

Please find below my responses to specific questions. Most of these are generic, or tie in to the PDC consultation. I also attach a blog post on the PDC, which puts the ideas in a slightly more cohesive order than answered in your consultation.

For the last decade, I have worked within academia on access to Government microdata, including UK microdata from the 2001 census (where different disclosure policies in the differnt countries had to be accounted for). I have also been an active member of the Open Data community since long before it was called that; I also run the OpenTech conference which data.gov.uk sponsored in 2010. All my comments are in my own personal capacity, blah blah blah.

Original link to post: www.disruptiveproactivity.com/?p=488

Short answers

1. Strengthening the OPSI directive, and potentially giving them a small, transparent, pot of money, from which small upfront costs could be requested to do initial work (SDC or technical) in return for the ongoing data supply to be fully-OGL. If not, that money should be returned to the pot.

3. In the spirit of transparency, all of the relevant metrics should be published by the organisations, with central Government bodies (ie bodies for which there are organograms) being available in some format as described, with soft-power incentives for others to follow, and a suggestion that an external organisation produces some form of credible and independent ranking. The Guardian may be an obvious example given their existing work in this area.

2 & 5. The publication of the data inventory would be critical, and would allow people to fin what data was available. A requirement on Responsible Owners of Data inside Government to ensure that their record in that database was present and correct, would aid it's ongoing accuracy. In addition, all publicly funded data holders or archives should be expected to produce something similar, to the same format, for the entire Public holdings to be catalogued, not just central Governments. This should be done "voluntarily", with an expectation that it will be a requirement in the granting of any future public funds approved by Publicly Funded money granting bodies. Additionally, Places of Deposit accredited by The National Archives are not currently required to publish full and machine readable catalogues as part of their accreditation; on any renewal, this should change, and standards should be maintained as appropriate.

The UK has a number of world leading Archives; if only people could find what was in them without going to that particular silo. Some are great; some, well, need some encouragement.

6. While the growth agenda drives a significant amount of Goverment policy, Open Data has additional scope. Knowing that a bus is 2 minutes late and you can walk for it, not run, will probably have some financial benefits (lower costs to the NHS of broken legs from falls for example), the real and substantive benefits are to people's lives and their daily experience. A known 5 minute wait meaning I can grab a muffin and mocha, but outweighed from the knowledge that has replace uncertainty. That is not measurable, easily, for any specific dataset, but should not be ignored.

The Two Consultations

The PDC is not about open data; but other data. I acknowledge and appreciate the differences. However, the Cabinet Office, in consultation events, publicly say that that this group is nothing to do with the PDC; the PDC people say they are working closely with the Cabinet Office. Both can be true; but there's clearly a lack of connected messaging.

Such disconnect is fundamentally damaging to the potential for "Making Open Data Real" as there are open datasets that can and must be derived from non-open data sources. That engagement needs to be central and balanced.There are and will continue to be ongoing tensions between access, use, privacy and cost. They pull in their own directions, in different ways depending on different details. Attempting to hide those tensions is a courageous decision, in a ways that could potentially suggest more than a "minor error of judgement and lack of oversight". While I appreciate that your communications budget has been cut, I'm not entirely sure that is the sort of "free publicity" that would benefit the agenda.

While starting the data release with the NPD, a compulsory, comprehensive, longitudinal, valuable dataset of every state-school child in the country, is a decision on which a rightful amount of focus will be given; and the principles there should generalise, there needs to be consideration given, independently.

The level of reviews, privacy considerations, etc, given to the NPD release is good, however, with little detail on what will be released as yet, there are likely to be a number of disagreements from the divergent assumptions based on the lack of information provided so far. I doubt you'll release pupil level microdata; or the raw text field for ethnicity, which, typos and all, would probably equally be a bad idea. In term of potential outputs in that case, a dataset of "real" catchment areas for schools (rather than supposed) could easily be derived to be open data, after simple SDC, as it includes no data about children at all. That could (and arguably should) be Open Data; whereas most datasets should be much more restricted. Those restrictions should not only be thought of as financial, but can be zero-cost with other restrictions.

There is still a large amount of grey areas in the middle of this dataset, and the many others that will be considered for release in future. How they get filled in, needs equally deep consideration. For the NPD, it may get it; for others, possibly not so much. For Open Data to be Real into the future, they must do so. Unlike many other policy areas, with open data, once data is out, the OGL is unrevokable on a particular dataset. Getting it wrong is a permanent state; and data has a zero cost of copying. The Treasury team got very lucky with the very first edition of OGL COINS as it was a tightly restricted release, to people who all obeyed the restrictions.

Given the murmurings and detail of the official denials around Health data and ONS microdata being made available under OGL or via the PDC, there should be very tight scrutiny of decisions around not just what is Open, but what is released. As licensing becomes less restrictive, protection of data, in terms of granularity, must go up, and, as a direct result, utility goes down. As an analogy to transport data, if the fully-open live bus information was only in 5 minute buckets, that would not be that useful to anyone.