Written Evidence Submitted by Owen Boswarva to PASC Inquiry on Statistics and Open Data
3 September 2013

Summary

- Open data has considerable potential to make the information economy fairer and more efficient.
- The Government's open data programme has given too much priority to the transparency agenda and not enough to release of useful "core reference" datasets.
- Progress has been slow and few public authorities have committed to significant releases of open data.
- There is no ingrained opposition to open data within the civil service but departments need clearer direction from the executive level.
- The Cabinet Office's approach to monitoring progress on open data is muddled and has little relevance to real-world priorities.
- The Department for Transport stands out as an exception and the rest of Government would do well to emulate its approach to open data.
- The Office for Public Sector Information has an inexplicably low profile.

Note

1. I am a data consultant and open data activist, and also represent the interests of external data users as a non-executive member of the Defra Network Transparency Panel. My professional background is mainly in modelling of risk data for the insurance industry. I am submitting evidence to this inquiry in a personal capacity.

Questions

Q1. Why is open data important?

2. One of the curiosities of the UK open data movement is that the broad principle, that publicly-funded datasets should be available for re-use at marginal cost under an open licence, enjoys support from people with viewpoints across the political spectrum. Where we tend to disagree is on priorities and practical implementation.

3. My own view is that open data has considerable potential to create fairer markets, by removing information asymmetries and increasing liquidity so that participants can negotiate transactions on a more equal footing. Charging for publicly owned datasets, particularly when the data holder has a monopoly, tends to skew the markets that rely on that data in favour of larger participants that can most easily absorb licensing fees. Data charges are an input cost and create barriers to entry for SMEs. Open data licences remove those barriers, as well as opening up additional modes of re-use (on the web especially) that cannot be accommodated by fee-based licensing arrangements. Even within large companies, use of open data promotes more efficient working practices because the data can be shared in a "frictionless" manner with suppliers and clients.

4. I am less convinced by arguments for open data as a driver for better government, or by the notion that "armchair auditors" can replace expert oversight of public finances. The increased availability of spending and performance data in a re-usable format is welcome, but no substitute for meaningful public consultation and open decision-making. My main criticism of the current Government's open data policy is that it has been subsumed into a rather amorphous "transparency" agenda. Government should do more to discourage commercial licensing of public sector datasets, and concentrate on releasing economically useful "core reference data" for open re-use.

Q2. Why does the Government need an open data strategy?

5. I am not convinced the Government does need an open data strategy in the form currently conceived, except perhaps for purposes of presentation. Over the past few years we have had ministerial letters, a white paper, a review, consultations, at least two statements of principle, a dozen sector boards have been set up -- yet when it comes to release of the most valuable public data assets the Government seems afraid to commit. My impression is that within the rank and file of public services there is no particular resistance to open release of data; the problem is a lack of any clear go-ahead from ministers and senior civil servants.

Q3. What should the Government's aims be for the release of open data?
a. Are the Government's stated key outcomes in its Open Data Strategy the right ones?

6. At the risk of stating the obvious, the Government's primary aim for the release of open data should be to actually release as much public data as possible for re-use under an open licence. The key prerequisite activity is to identify all of the datasets held by government, prioritise them based on their potential for beneficial re-use, and overcome practical barriers to their release (where release is economically feasible and consistent with other imperatives such as data protection and security).

7. Unfortunately the indiscriminate dumping of small, low-value datasets on Data.gov.uk has created the illusion of progress -- 9,000 datasets sounds like a lot, but what proportion is that of what total? The Cabinet Office's very recent attempt to produce an inventory of "National Information Infrastructure" is on the right track, but it is too early to tell whether that initiative will generate any wholesale release of open data. (I am mindful that the previous Government made a similar attempt with its Information Asset Registers.)

8. I am unsure what the Committee means by "the Government's stated key outcomes in its Open Data Strategy". There are departmental Open Data Strategies but I not aware of a cross-government version other than the Open Data White Paper, which does not seem to list key outcomes. If this is a reference to the eleven "Corporate Commitments" as tabled in the most recent Written Ministerial Statement, I would say that those are of limited relevance to the performance of the Government's open data strategy as a whole. Public data with potential for re-use is not evenly distributed across Government, so metrics that compare one

department to another are not very meaningful. In my view however the public bodies making the greatest contribution to open data in the UK are ONS (as an inherent part of its role) and the Department for Transport (because it has simply decided to get the job done).

Q4. How can those engaged in open data, and those engaged in producing government statistics work together effectively to produce new data?

9. Strictly speaking, as I understand it, the production of new data is not part of the open data agenda. Open data is typically about unlocking public datasets that already exist so that they can be re-used for secondary purposes. The rationale is that these datasets should be open because they have already been funded to deliver a public task, so arguing for production of new data doesn't really sit within the open data agenda itself. In practice of course there are many people with an interest both in the production of data to deliver a public task and re-use of that data for secondary purposes, since those two categories of use are often complementary.

10. However, on this theme generally, it would be very desirable to have more involvement from government statisticians in the identification and prioritisation of public data for open release. Some supporters of open data tend to treat statistics as if they are not "real" datasets, but that is not my view. (On the contrary I regard small area Census outputs as the nation's single most valuable open data asset.) The Government has engaged extensively with IT developers and with entrepreneurs in the "app economy", but has somewhat neglected the potential for modelling and derivation of analytic insights from open data; this is an area where input from ONS staff would be useful.

Q5. How can more statistics and administrative data of all kinds become more freely available?

11. Statistics are usually produced with open release in mind, and there are few of the barriers to open release that we find with the underlying administrative data or with publicly-funded scientific and technical datasets. From a re-use perspective the main difficulty with statistics is the frequency with which series are cancelled or redesigned to meet new government priorities or reporting requirements. This is a particular issue under the current Government; some quite significant statistical datasets have fallen to the axes of austerity, "red tape" or ministerial ill favour.

12. With respect to administrative data it is difficult to generalise, but Government could do more to highlight the data sources maintained by departments; many remain quite obscure, though every government department does make an official statement of administrative sources to the UK Statistics Authority. Some administrative datasets contain personal data, and I know some open data advocates would like to redraw the lines of data protection to make more of that data available for re-use in an "anonymised" form. My own view is that protection of personal data should remain paramount.

Q6. Is open data presented well and of adequate quality?
a. Are the formats of the data being published accessible, useable and

understandable to the public?
b. What metadata is needed to make releases useful?
c. Who will use the data released?

13. It is impossible to make a general statement about the presentation and quality of open data. That depends on the specifics of individual datasets. There is considerable scope for variation, just as there is with closed data. My priority is release of the data under an open licence. Basic accessibility is a necessity, but beyond that I would rather public authorities released their bulk data quickly than worry too much about formats, cleansing, how many stars the data has, etc.

14. There is a tendency among some users to complain if an open dataset has not been "cut" to their idea of an ideal specification, but those complaints are often unreasonable. Public authorities cannot anticipate all the potential ways in which their data might be re-used, and should not be expected to try. Users should expect to carry out their own manipulation of the data in order to make it suitable for their purposes. (I am referring here to secondary use of the data, i.e. re-use, not to a public authority's presentation of their own data to their own end users.)

15. Metadata standards are well understood (the Data.gov.uk pro forma is a reasonable example); the real question is how much documentation a public authority should provide with an open data release. It is difficult to argue that data should be released without at least adequate contextual information, and that may involve writing new material. Even if there is internal documentation already available, it may assume some level of institutional knowledge not available to external users.

16. The Committee's question as to who will use the data released is difficult to parse. There is no general answer; the potential for re-use of an open dataset depends on the characteristics of the individual dataset. Given that open data licensing makes it much easier to use data on the web and in mobile applications I suppose we can say that, all else being equal, an open dataset is more likely to reach a wider range of users than would a closed dataset with otherwise similar characteristics.

Q7. How successful has the Government's Open Data initiative been in changing behaviour in the Civil Service and wider public sector?

17. This may be an unfashionable view, but as far as I can tell there is no ingrained opposition to open data within the Civil Service or the wider public sector. To the extent that there has been any cultural change or improvement in understanding, it is difficult to disaggregate the effects of the Government's open data policies from the broader influence of the open data movement in civil society. My main observation is that there is a disconnect between the Cabinet Office's rhetoric on open data and practical implementation by key delivery departments. The Ministry of Justice has successfully fought off European proposals to strengthen the PSI Directive, and given UK public authorities new powers to charge for re-use of data under FOI. BIS has so far protected most of the "crown jewels" of public data from open data release (i.e. the Royal Mail's Postcode Address File and the assets of the Public Data Group trading funds). Other commercial licensing operations, such as those at the Environment Agency and the British

Geological Survey, remain unchallenged. If there is a need for behavioural change, it is most likely at the policy-making level.

Q8. Which datasets are the most important?
a. What are the best examples of data being made open and resultant benefits to business or society?

18. The concept of prioritising core reference datasets is sound, even if I have misgivings about the approach the Cabinet Office is taking to identify those datasets. As a rule of thumb, the datasets with most potential for re-use as open data are the datasets Government departments (or more often executive agencies) are most reluctant to release -- because their economic value has already been recognised and they are using them to generate licensing revenue.

19. I would put the following at the top of my list of important datasets: national address data (held by Royal Mail and Ordnance Survey), large scale national mapping and aerial imagery (Ordnance Survey), national cadastral polygons (Ordnance Survey and Land Registry), flood risk mapping (Environment Agency, Sepa, local government), historical weather observations (Met Office).

20. As best examples, I would nominate:

    - small area Census outputs and by extension the IMD (for their pervasive contribution to sociodemographic research and allocation of public resources),
    - Ordnance Survey's OpenData products (projected to generate net growth in GDP of between £13.0m and £28.5m per annum by 2016 according to an economic value study prepared last year),
    - prescribing data by GP practice from the HSCIC (for its analytic potential),
    - the wide range of transport data available from DfT and the rail operating companies (for kick-starting interest in the potential for open data re-use in mobile apps), and
    - data.police.uk (for regularising the availability of consistent crime incident data).

21. The question the Committee should consider, however, is how many significant open data releases are traceable to the Government's central programme or strategy. I have included four examples from the current period of Government, but it is difficult to see the influence of the Cabinet Office in any of them. A few smaller open data releases did emerge eventually from the measures announced in the 2011 Autumn Statement (Land Registry's Price Paid Data, basic company data from Companies House, etc.). However the Cabinet Office's formal demand-led pipeline for unlocking datasets, based on a budget given to the Data Strategy Board to "buy back" public data, has apparently not produced any outputs in its first year of operation.

Q9. How effective is the work being undertaken by the Cabinet Office to monitor the progress of Departments in publishing their agreed datasets?

22. The Open Data Strategies released by each department in June 2012 set a low bar for expectations. Most of the strategies followed a template, with a substantial

amount of commentary but few commitments to release specific datasets as open data. (DCLG's document ran to 39 pages and noted on page 27 that the department had no plans to release any new datasets during the strategy period.)

23. A member of the Cabinet Office's Transparency Team did collate a very helpful spreadsheet of datasets mentioned in the departmental strategies. However rather than take that as the logical baseline to monitor progress, the Cabinet Office seems to have decided to try a smorgasbord of approaches. In the first WMS report we had "openness scoring", apparently driven mainly by the format of each department's dataset records on Data.gov.uk. In the next two WMS reports we had the table of "corporate commitments" apparently selected only because they were common across departments. More recently the Cabinet Office has told us that Data.gov.uk is developing new tools to measure and compare department's adherence to commitments, and that once that functionality is bedded in the WMS returns will be stopped. None of this seems to have very much to do with measuring the extent to which departments are contributing to the open data agenda in real-world terms.

Additional Comments

24. These are several further points that the Committee may wish to consider as recommendations, as I think they would help to improve the Government's approach to open data or at least reduce some weaknesses in that approach.

25. The terms of reference given to the Data Strategy Board, and by extension to the Open Data User Group, require potential users to make a business case for open release of public data. This is contrary to the principle of "Open Data by Default", endorsed by the Government in the Public Data Principles and in the G8 Charter. If a dataset is closed, the onus in the first instance should be on the public authority to make a business case for it to remain closed; to demonstrate either that there is some insurmountable barrier to release (such as data protection or national security) or that the dataset is already available on commercial terms that do not discourage re-use. As a corollary to that approach, public authorities should be required to disclose any income they currently receive from licensing of data so that we can judge whether an open data approach would be more or less beneficial.

26. The Cabinet Office should collate and publish a list of all the departmental transparency “sector boards” now in operation, with the members of each, and require them to publish minutes of their meetings within a reasonable timeframe.

27. The Office of Public Sector Information has no discernable presence on the web. Since the opsi.gov.uk domain was retired in 2010, all material produced by OPSI has been buried in an undifferentiated area of the National Archives website. OPSI remains the UK's official PSI regulator, and members of the public are supposed to be able to make complaints to OPSI on matters within its remit. At minimum OPSI should have its own landing page on the National Archives website, so that the public can find it easily via Google or GOV.UK.

28. Parliament is necessarily outside the scope of the Government's open data programme. However I would urge the Committee to consider engaging with the open data community separately on bulk release of datasets held by Parliament. There are several datasets, such as MPs' contact details and the various Registers of Interest, that Parliament could perhaps make more readily available for re-use in the interests of accountability and democratic transparency.

29. I thank the Committee for the opportunity to submit evidence to this inquiry.