RESPONSE TO CABINET OFFICE OPEN DATA CONSULTATION

Kieron O'Hara

Electronics and Computer Science
University of Southampton
Highfield
Southampton SO17 1BJ

Dear Sir or Madam,

I would like to respond to the Cabinet Office's Open Data Consultation, based on the document currently available at http://www.cabinetoffice.gov.uk/sites/default/files/resources/Open-Data-Consultation.pdf. In general, I am very supportive of the transparency programme, which I believe has the potential to transform government for the better by empowering citizens.

My report for the Cabinet Office, *Transparent Government, Not Transparent Citizens*, has already discussed many issues related to transparency at some length, and I hope that the 14 recommendations in that report will be considered in the composition of the White Paper. It is not my aim to restate these recommendations in detail. However, as the consultation document makes very little mention of privacy, I do want to add a few words on the topic. Most of my comments can be gathered together under the headings proposed by the consultation document.

ENHANCED RIGHT TO DATA

1. The consultation document is correct to identify the importance of enhancing rights to data and in changing culture in public bodies. The existence of a right to data is an important addition to the utilitarian arguments for open data and transparency (see section 7 and the annex of the consultation document). If takeup of open data is not as great as hoped, then the utilitarian benefits might be slower to emerge, which might prompt calls for the programme to be scaled back. In contrast, the rights-based argument, which is founded on legitimacy, is unaffected by slow takeup.

This is important because the demand for open data is currently not fully understood, and in any case one of the points about transparency is that value-adding uses of open data by innovative entrepreneurs are extremely hard to predict or anticipate. One would naturally expect open data to exhibit a 'long tail' – some datasets being in common use, while many others, perhaps a majority, used only rarely for niche purposes. The value of a comprehensive open data programme is making as large a suite of information as possible available to information entrepreneurs to devise innovative services.

2. The relation between open data and FoIA needs to be fleshed out in more detail – in particular to look at the role of FoIA in a world where citizens had a right to data. This matters not only conceptually, but also in consideration of cost. It would be much cheaper to publish information routinely, than to undertake the FoI procedure every time there was a request for information. As the document states, ICT management is important; there are several tools for routinely and instantly publishing spreadsheets and other documents on the Web. In the longer term, use of W3C open standards for representation would enhance the value of data for citizens. Apart from the one-

off costs of ICT upgrades (which will happen regularly anyway), such measures should dramatically reduce information handling and publishing costs.

There are several possible relationships between open data and FoI. For example, one might make the strong assumption that if information is FoI-able (i.e. if it *could* appear in the public domain), then it *should* appear as open data. In that case, FoI would take the *de facto* role of an appeal against a decision not to treat some data as open data. A weaker assumption about open data would leave a wider role for FoI. But however open data were finally defined, one would hope (a) that the number of FoI requests should fall, and (b) that the costs associated with FoI would fall by a proportionately greater amount, as economies of scale in publication were realised.

3. Specifically, I see no especial need to create a new independent body, nor to expand the powers of the Information Commissioner's Office at this stage.

4. With respect to the privacy considerations in question 3, privacy-protecting measures should be kept under review. When the default is not to publish, as now, privacy protection is enhanced. If the culture is changed so that the default is to publish, accidental privacy breaches become more likely. Hence under a right-to-data regime, privacy protection would need to be kept under review.

CORPORATE RESPONSIBILITY

1. All sectors would benefit from smaller-scale versions of the Transparency Board, to bring experts and stakeholders together to assess demand for data, and to consider those issues (such as privacy or national security) where extra care is needed to weigh costs and benefits of publication.

STANDARDS

1. The current set of Public Data Principles, and Berners-Lee's 5* rating system, are a sensible basis for open data, and should remain the pillars of the transparency programme.

In particular, the Public Data Principles' championing of making data available for free is important. Selling data will inevitably result in information monopolies, leading to an information market characterised by rent-seeking, not value-adding.

That is not to say that data should never be sold; my report agrees that there are circumstances where applying terms and conditions, or identification and registration, or even charging, will be sensible. But this should be the exception. The main point is that innovative services will be created when there is equal access to data for all potential service providers. There is no problem with monetising innovative *services* where possible, perhaps through subscriptions or charging. In that case, providers would be incentivised to be innovative, as long as free access to data ensured competition.

2. Accrediting information intermediaries would be detrimental. Government may have a role in encouraging the appearance of intermediaries, but part of the point of transparency is for the government to 'let go'. If government favours certain intermediaries with 'kite marks', then the market for information services based on open data would inevitably be distorted. Intermediaries would be incentivised to meet the criteria for accreditation, rather than to provide innovative services.

Having said that, there *is* a role for government to play as an intermediary itself. As with police.uk, the highly successful Home Office crime mapping site, government agencies and departments could present the data to citizens in order to develop the constituency of open data users. Police.uk has helped introduce people to crime mapping, and one could imagine sites funded by other government agencies and departments meeting a similar function (e.g. in education or health). Of course, the agency should simultaneously release the data, so that other intermediaries could supply supplementary services. If the government opts for an intermediary role, it should ensure that by doing so it does not raise the barriers to entry to the information market in that sector.

MEANINGFUL OPEN DATA

1. Issues of quality are best addressed by transparency. Flaws in datasets are more likely to be spotted if they are used more widely, and if the demand-side has an influential enough voice, government agencies and departments would be driven to improve quality. Furthermore, benchmarking and peer group pressure across agencies and departments will also have a role to play.

2. In fact, the reaction from the demand side would be an important part of the definition of 'quality' canvassed in question 5.

3. It should also be pointed out that if data are of such poor quality that the agency or department is embarrassed to release them, then they shouldn't be used by the government either. It would not be cheering to hear that some government services rely on data which are too low in quality to release to the public. No department could make this admission and retain a reputation for quality of service.

GOVERNMENT SETTING AN EXAMPLE

1. It does not matter where data are stored, although departmental portals could be used to give meaningful context to data releases. It is extremely valuable to ensure that all public open data can be accessed from data.gov.uk, as one (but not the only) access point. The other major point is to ensure that datasets can be easily found via search engines.

2. There are a number of issues raised in the consultation document about prioritisation. These questions are very hard to answer in the abstract. Broadly speaking, it would be sensible to experiment with different methods in different sectors. But ultimately, prioritisation is best achieved by listening to demand and releasing the information people ask for.

3. Similarly, the question whether to release broad or deep data is open. My instinct would be to go broad, to cover as many sectors as possible. Intermediaries in those sectors would then – given relatively non-detailed data – be able to make a more informed estimate of what detailed data would meet their purposes.

The converse strategy, releasing data in depth, would undertake a greater risk of releasing unnecessarily detailed datasets (i.e. datasets whose detail would not add to their utility), while also failing to build up demand in those sectors which were not covered at all by data releases.

INNOVATION WITH OPEN DATA

1. Yes, government can stimulate innovation, and as discussed earlier, one method is for government to play the role of intermediary in some sectors.

Another method, again discussed above, is for government to listen and react to demand for data.

Thirdly, once more mentioned above, the incentives for innovation will be maximised if (a) public open data to act as input to services was available without restriction to all service providers in the market, and (b) service providers were allowed to monetise services that add social and economic value to the data. Conversely, charging for data will tend to promote rent-seeking.

A fourth method, already undertaken, is for innovative apps to be showcased on data.gov.uk. The existence of good examples of information-based services will drive both the demand for and the supply of further services.

PRIVACY

1. When writing my report, I was specifically asked to address the question of jigsaw identification of citizens through public open data.

Although I found that the risk of this is small, it is a real risk in the context of government releases of anonymised datasets. It also requires empirical investigation to quantify that risk.

It is true to say that anonymised data have been shared for years without serious incident. This is an important ground for optimism.

However, it does not take into account the *cumulative* nature of the risk, which increases (a) with the quantity of relevant data on the Web, and (b) available computing power. Neither does it take into account the fact that most data sharing in the past has taken place in controlled, managed conditions. This is not true of data downloaded from data.gov.uk.

It is also a *non sequitur* to argue (correctly) that it is unclear how anyone would gain from jigsaw identification of UK citizens, in the context of recent research (reported for instance in the October 2011 edition of *Scientific American*, p.76) that "the truly enormous [data] breaches have increasingly been carried out by 'hacktivists' – individuals or groups who are angry about an organization's actions." Such hacktivists achieve no personal gain from their attacks, and it is worth remembering that of all organisations, government is perhaps one of those with a tendency to attract anger.

If the transparency programme is seen to ignore an unquantified risk, it is in danger of losing the confidence of citizens, or of incurring entirely avoidable criticism from civil liberties campaigners. Preserving public confidence is paramount for the long term future of transparency in the UK.

It follows from this that a programme of research into jigsaw identification would be an inexpensive way of reassuring the public, addressing the concerns of civil liberties campaigners, and protecting privacy.

2. More broadly, confidence in the transparency programme would be strengthened if privacy were built in, rather than bolted on. The notion that you can prepare data for release, and then assess privacy implications accurately before ‘pressing the publish button’ is false; it risks error.