Making Open Data Real: A Public Consultation

A response from Marion Oswald[1]

The author welcomes the opportunity to respond to the Government's consultation regarding its Open Data proposals. This response considers a number of questions from the consultation document, focussing particularly on the legal aspects and the wider issues regarding the overlap of freedom of information, re-use and copyright.

1 Glossary of key terms

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

The terms 'dataset' and 'data' are used interchangeably throughout the consultation document with some sections, including a number of the examples of best practice in Annex 1, referring to transparency of 'information.' On the assumption that it would be the ultimate aim to transpose the new rights and principles discussed in the consultation into law, clarity as to the scope of public sector responsibilities will be crucial, both from a compliance and an enforceability perspective. The following may be areas where further consideration would beneficial:

1.1.1 The term 'dataset' is described in the glossary by reference to the term 'data,' structured or unstructured, including datasets 'about' public services. There are certain differences between this description and the definition of dataset used in clause 100(2)(c) of the latest Protection of Freedoms Bill,[2] which limits its definition to 'raw' information and excludes organised, adapted or otherwise altered information (i.e. value-added information) from the proposed disclosure and re-use obligations. By use of the words 'structured' and 'about' public services, the terms in the consultation potentially go further than envisaged by the Bill, in that the consultation's terms could be regarded as covering such value-added information.

1.1.2 It is stated that, for non-government bodies, information about aspects unrelated to the delivery of their 'public service function' are not in scope. Questions may arise in relation to public bodies such as Universities, and others which have hybrid functions. Would information related to research be unrelated to a University's public function for instance? Freedom of information decisions have decided that research data does fall within the scope of the Freedom of Information Act (the FOIA).[3] HE institutions have expressed concern regarding the potential long-term implications of such decisions and during the report stage of the Protection of Freedoms Bill, a new clause was moved making information obtained in the course of, or derived from, a programme of research exempt information for the purposes of the FOIA, subject to a number of provisos. As the Open Data proposals develop, will the

[1] Solicitor, Senior Lecturer, Department of Law (Centre for Information Rights), University of Winchester, marion.oswald@winchester.ac.uk

current FOIA exemptions require enhancement or refinement in the context of proactive, rather than reactive, disclosure requirements?

1.1.3 The Protection of Freedoms Bill is perhaps not the only example of parallel and overlapping initiatives in the field of public sector data transparency. The Information Commissioner has recently announced a consultation on the content of publication schemes;[4] the Department for Communities and Local Government has published a Code of Recommended Practice for council transparency;[5] the FOIA will soon be subject to post-legislative scrutiny. The current consultation process appears to represent an ideal opportunity to initiate a comprehensive and coordinated review of the various legal regimes governing both access and re-use of public sector information. Without this, there is surely a risk that piecemeal reforms may create further confusion for public sector bodies and recipients of data, thus increasing the likelihood of legal challenge and increased public costs. It is suggested that if the Government were to consider rolling out its Open Data proposals to the entire public sector, then a transitional approach may be appropriate in respect of applicability to particular bodies/sectors and the type of information covered.

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

Which datasets would be in scope? The implications of the distinction between 'information' and 'data' may be relevant to this question. Kieron O'Hara said that the distinction 'is more often gestured towards than defined rigorously, but broadly speaking data are at a lower level of abstraction than information. Information is data interpreted for some audience in some way.'[6] His report then concentrated on potential privacy threats from data transparency, rather than information transparency more generally.

In terms of the Government's Open Data aims however, it must be clear which 'information' or 'data' are covered by any new obligations. A review of current legislation indicates that

---

[2] A dataset is defined as a collection of information held in electronic form where all or most of the information in the collection (a) has been obtained or recorded for the purpose of providing a public authority with information in connection with the provision of a service by the authority or the carrying out of any other function of the authority, (b) is factual information which—(i) is not the product of analysis or interpretation other than calculation, and (ii) is not an official statistic, and (c) remains presented in a way that (except for the purpose of forming part of the collection) has not been organised, adapted or otherwise materially altered since it was obtained or recorded

[3] For example, Information Commissioner Decision Notice FS50163282, 29 March 2010, in relation to tree ring research data held by Queen's University Belfast

[4] http://www.ico.gov.uk/about_us/consultations/our_consultations.aspx

[5] http://www.communities.gov.uk/publications/localgovernment/transparencycode

[6] Kieron O'Hara (2011) *Transparent Government, Not Transparent Citizens: A Report on Privacy and Transparency for the Cabinet Office*, para 1.3.4, p 14

the term 'data' has not necessarily been used to indicate 'a lower level of abstraction.' Section 1(1) of the Data Protection Act 1998 states that 'data' means information which are electronically processed or recorded in a relevant filing system i.e. structured so that information is readily available. Personal data can include expressions of opinion. The FOIA is concerned with 'information' recorded in any form.[7] The Re-Use of Public Sector Information Regulations deal with the re-use of 'documents'; defined in regulation 2 as 'any content, including any part of such content, whether in writing or stored in electronic form or as a sound, visual or audio-visual recording, other than a computer program.'

Of course datasets, whether structured or unstructured, are often useless or irrelevant without the accompanying descriptions of the dataset, such as data dictionaries, field/column descriptions, etc. It would be unhelpful to the Government's Open Data aims if such accompanying descriptions were to fall out of scope. There may be a dilemma though, in that for maximum availability, data will need to be published in a publicly accessible format, but the dataset may have been collected and stored in a proprietary format that requires the purchase of software to read it. If a public authority is required to re-process data into an open format, this will necessarily come with cost.

Unstructured data, such as emails and memos are, as the Information Commissioner has commented, 'important in delivering accountability.'[8] However, the very nature of unstructured data defies categorisation, thus increasing the difficulty of imposing proactive publication obligations in respect of it. If unstructured data is to fall within scope of the Open Data agenda, then emails and memos will become susceptible to proactive publication; the assessment of each one of these to ensure that FOIA exemptions are addressed (personal data, policy advice to ministers, commercially sensitive aspects etc) would be a substantial burden on any public authority.

So how could a test be determined? Rather than attempting to create a (sometimes rather artificial) distinction between 'data' and 'information', perhaps a more effective approach would be to focus on a test that deals with a) how the information has been collected; b) for whom the information is collected; c) the purpose(s) of the information; d) how the information is held e.g. in a structured and publicly available format; e) whether the information would require re-formatting f) whether the information falls within any excluded categories? This approach is likely to require a detailed sectorial approach to be taken but with the aim that greater clarity would be achieved in the long term.

In terms of excluded categories, these are likely to depend on the extent to which the information can be re-used without restriction, or whether any intermediate options will be

[7] s 84

[8] http://www.ico.gov.uk/news/latest_news/2011/the-citizen-should-be-in-the-driving-seat-over-what-information-is-made-public-28092011.aspx

available, such as that permitted by the Non-Commercial Government Licence. Point 2.1 below also refers.

2 An enhanced right to data

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

Page 22 of the consultation document asked how it would be possible to establish stronger rights for individuals, businesses and other actors to 'obtain, use and re-use' public data. The obtaining and subsequent re-use of data are often discussed in the same breath. To date, however, separate (albeit linked) legal regimes have governed access and re-use of public sector information: access by the FOIA and the Environmental Information Regulations, and re-use by the PSI Regulations. The Protection of Freedoms Bill contains mandatory re-use obligations in respect of 'datasets' which would apply to all public bodies covered by the FOIA, including those institutions such as universities, libraries and museums that are currently excluded from the ambit of the PSI Regulations. This would result in two re-use regimes, one mandatory, one not, and two different regulators: a position far from ideal.

Provided it was possible to define clearly the categories of data to which the obligations applied, enhancing the content of publication schemes would seem to be the most straightforward way to strengthen the obligation to *publish* proactively. Dealing with *re-use* is more complicated, particularly where copyright works, commercial interests or third party rights are in issue. Recent cases have shown some public authorities to be willing to use copyright arguments to prevent further publication of disclosed information on websites such as http://www.whatdotheyknow.com/ and other online media, even though such online publication would have had little or no impact on the authority's interests.[9] The Information Commissioner has not yet stepped fully into the online publication debate, on the basis that re-use matters are beyond the jurisdiction of the FOIA.

Conversely, it seems unrealistic to trust that providing access to a copyright work under the FOIA via electronic means will not inevitably result in electronic distribution. For some public authorities or data, this may have little or no adverse consequences, and such distribution may well be encouraged. Those public authorities with interests linked to the commercial sector or with long-term research interests may well be concerned about a long-term prejudicial effect, with use of the Open Government Licence (or even the Non-Commercial Licence) providing minimal comfort.

The time seems to be right to step back and consider the long-term consequences of the gradual merging of re-use obligation into information access regimes, and to ensure that any developments take steps towards a more joined-up approach. In addition, the following suggestions are put forward for consideration:

[9] See for example Information Commissioner Decision Notice, FS50296350, 2nd December 2010, regarding Brent Borough Council

2.1.1 A new copyright exception which would allow public sector copyright works released under the FOIA to be further published online, provided that such use does not substantially prejudice the owner's interests, or is not 'unfair.'[10] Whether substantial prejudice or fairness concepts could appropriately deal with the diverse interests of UK public authorities, including their relationships with third party copyright owners, would need further review. Commercial versus non-commercial definitions may not be helpful; is university research 'commercial' for example? Should such an exemption be limited to the purpose of communication, review and debate of the results of the FOIA request? To be effective such an exception should not be able to be overridden by contract and would need to incorporate an element of future-proofing against inevitable technological developments.

2.1.2 In the United States, copyright protection is not afforded to any work of the United States Federal Government, although the Government is not prevented from receiving and holding copyrights transferred to it.[11] It may well be possible to identify categories of UK public sector information to which a copyright-free status would be appropriate (and perhaps those datasets proactively released by the Government are an example), thus avoiding copyright licensing concerns when such information is released under the FOIA or proactively disclosed. FOIA exemptions should of course remain to protect information not appropriate for disclosure. There are of course likely to be considerable complexities to be considered in relation to works created in combination with third parties or with private sector involvement.

2.1.3 The Government's Open Data agenda would appear to lend itself to the development of a set of 'principles' that, in a similar way to the Data Protection Principles, could then form the backbone of any new law. Those principles could apply to the information that is to be released proactively and subject to re-use obligations. The following rough drafts are put forward for further debate: '[the defined information] shall be published electronically as soon as reasonably practical after it becomes available; [the defined information] shall be updated on a regular basis; [the defined information] shall be released in accordance with appropriate open standards.'

2.2 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?

The author would support O'Hara's recommendations 3 and 9 in relation to the role of the Information Commissioner.

[10] Chris Reed has advocated a restructuring of copyright law so that it gives creators appropriate control over the 'use' of their works. *Open Culture – Rethinking the Legal Framework*, Computers & Law Magazine Vol. 21, Issue 6, page 19

[11] Title 17, U.S.C. s105 at http://www.copyright.gov/title17/92chap1.html#105

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

As explored extensively in the O'Hara report, a risk exists of the efficacy of the anonymisation of datasets being undermined if the same anonymisation method was used in multiple datasets or if datasets can be cross-referenced with other information or knowledge. It is perhaps rather overstating the position to say that the definition of personal data in the Data Protection Act 1998 allows 'rampant reidentification.'[12] The definition in s1(1)[13] serves to identify categories of data to which the Act's safeguards attach, limb B of the definition requiring an assessment of whether identification could occur by possession of the anonymised data and other 'information.'

Bearing in mind the significant steps needed to achieve de-anonymisation, and the difficulty of allocating responsibility for 'jigsaw' identification, the perspective of the data controller in identifying personal data surely remains key. A case-by-case approach should continue to be taken to the assessment of datasets for release; recent FOIA decisions regarding the personal data exemption illustrate the importance of such an approach, in particular regarding the effectiveness of anonymisation.

For instance, the withholding of information about the numbers of teaching staff who had been investigated for sexual offences was upheld by the First-tier Tribunal (Information Rights) on the basis that other information existed in the public domain that meant that there was a strong likelihood of identification taking place. Therefore the information requested was 'personal data' to which the DPA applied.[14] In contrast, the High Court held that abortion statistics were not personal data.[15] It overturned the Tribunal which had held that statistical information ceased to be personal data only where it could no longer be cross-referenced to other information held by the data controller but that release was justified because there was insufficient risk of identification (following Baroness Hale in CSA[16]). Cranston J adopted Lord Hope's reasoning in CSA: that even though the data controller held the means to identify the individuals, this did 'not disable it from processing [the data] in such a way, consistently with recital 26 of the Directive, that it becomes data from which a living individual can no longer be identified.'[17] It was for the ICO to make a decision of fact

[12] Note 6, para 4.2.4, p 45

[13] Personal data means data which relate to a living individual who can be identified: (a) from those data, or (b) from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller.

[14] Smith v Information Commissioner and another [2011] EA/2011/0006

[15] R (Department of Health) v Information Commissioner [2011] EWHC 1430 (Admin)

[16] Common Services Agency v Scottish Information Commissioner [2008] UKHL 47, [2008] 1 WLR 155

[17] Note 16, para 27

as to whether any particular anonymisation technique would achieve this. Any other conclusion would, according to Cranston J, be 'divorced from reality.'[18]

There does however appear to be some legal uncertainty stemming from the judgment of Baroness Hale in CSA. In light of this and of technical developments, it would seem the right time for a review of the DPA definitions. As recommended by O'Hara, closer dialogue between legal and technical practitioners could usefully contribute to this.

2.4 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?

Please see point 1.2 above regarding unstructured data in particular.

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

Against the backdrop of the Government's recent tender notice for the creation of the G Cloud service, this response has considered 'cloud' services in particular. A provider of remote computing services is likely to be regarded as holding information on behalf of a public authority for the purposes of the FOIA, certainly not an area commonly covered in standard service terms and conditions.

From the public sector perspective, data availability and integrity will be crucial aspects of a cloud service: will the data be both available when it is needed and unchanged from the original? Standard terms often attempt to push ultimate responsibility for confidentiality and integrity onto the customer.[19] A cloud provider will need to be able to live up to a public authority's freedom of information commitments, including proactive disclosure and open format requirements.

The contract between a public authority and a provider of cloud services could of course be subject to disclosure, either proactively under the current transparency agenda, or in response to a FOIA request. But what if a provider uses the public authority's data, say to create a report: could it be argued that the report is held on behalf of the authority, even though it had a primary private purpose? It would be advisable for such potential issues to be anticipated by the contract.

Marion Oswald
Solicitor, Senior Lecturer
Department of Law
University of Winchester

[18] Note 15, para 54

[19] *Bradshaw S, Millard C, Walden I,* Contracts for clouds: comparison and analysis of the terms and conditions of cloud computing services*, I.J.L.&I.T. 2011, 19(3), 187-223, 202, para 4.7 and Bradshaw S, Millard C, Walden I,* Watching Cloud Contracts Take Shape*, Computers and Law, 7-10, 8-9*