Submission to the Cabinet Office Consultation: "Making Open Data Real"

EnCoRe project

20 January 2012

About EnCoRe

1) EnCoRe (http://www.encore-project.info) is a multi–disciplinary research project, spanning a number of IT and social science specialisms. It is researching how to improve the rigor and ease with which individuals can grant and, more importantly, revoke their consent to the use, storage and sharing of their personal data by others. Much of this data may form the underlying data sets for open data.

2) The project partners are Hewlett–Packard Laboratories, HW Communications, QinetiQ, the London School of Economics and the University of Oxford. The project runs from June 2008 to May 2012. It receives funding from the UK Government's Technology Strategy Board, Economic & Social Research Council and Engineering & Physical Sciences Research Council.

3) EnCoRe's overall vision is to make giving consent as reliable and easy as turning on a tap, and revoking that consent as reliable and easy as turning it off again. Turning this into reality, for both the individual and the organisation, requires

    - consent management technologies to be developed,
    - IT systems architectures that include these to be developed,
    - organisations' operational processes and systems to be designed or enhanced to use them,
    - easy-to-use interfaces to be developed and implemented, and
    - the regulatory regime that underpins all of this to be enhanced and strengthened.

4) EnCoRe is working on all of these areas to develop working technologies and model policies that allow the individual to take control over their own personal information. Such control could help maximise the sharing and re–use of such information whilst addressing privacy concerns. This becomes particularly relevant in the context of eGovernment, especially in Open Government and Government Cloud initiatives.

5) As a project EnCoRe has a particular interest in the potential privacy impacts of open data. Whilst it recognises that the consultation was issued before Kieron O'Hara's report on transparency and privacy had been issued, questions about privacy of personal data seem somewhat downplayed within the consultation. Thus, in response to question 3 (page 25) about whether existing safeguards are adequate to regulate the Open Data agenda, our response is simple: No.

6) Within the project, however, there is also expertise on the access, re–use and sharing of public sector information as well as data quality and this submission draws on this expertise as well.

EnCoRe's perspective

7) In the past five years we have seen a number of policies and regulatory instruments being introduced that aim at increasing access, re–use and sharing of public sector information. For example the PSI and INSPIRE EC Directives have emphasised the need to re–use and share public sector information while ensuring that data–protection rules are respected. While the relevant legislation implementing these directives repeats the commitment to data protection laws, there is a general lack of mechanisms (technical, organisational and regulatory) that allow the individual to control the flow of her personal data in shared– and open–data environments.

8) In many circumstances, open data is made available without the data subject's explicit and informed consent because it is assumed that existing techniques of anonymisation providing sufficient protections against re–identification and misuse of the underlying personal data. EnCoRe, however, has assessed the long–term potential risks of these techniques and has concerns about the long term viability of anonymisation used in these ways. Instead, it proposes that consent based mechanisms should be used such that when a data subject revokes the consent that they have given to a data controller to use their personal data for particular purposes, the data is put out of use. The current EnCoRe framework implements this through "sticky policies" that attach the consent preferences to the data they refer to.

9) Consent based alternatives to anonymisation, such as those provided by EnCoRe, provide a different perspective on the privacy issues of open data and the EnCoRe project provided evidence for Kieron O'Hara's privacy and transparency review.

10) As a result of this input, EnCoRe have been working with the Ministry of Justice and Department for Education about the potential risks associated with their plans to make personal data available in anonymised form. For example, we ran a small challenge whereby students worked with the MoJ to see if it would be possible re–identify individuals in the re–offender data sets that were released at the end of October 2011.

11) A second EnCoRe concern focuses on the relationship between privacy practices of data controllers and their approaches to data quality. Through a series of expert focus groups with a variety of organisations, a number of issues relating to the underlying quality of the data that might be disclosed through open data processes have been raised. This research suggests that overall data collection process should be based on the goal of quality—both the quality of the data collected at the source and the quality of data resulting from data processing activities.

12) Thus the key issues to be addressed in order to make data collection and management appropriate for ensuring high quality include:

- Clearly define the purposes and the uses of data collected. In order to collect data efficiently and to manage them effectively, an understanding of the reasons underlying data collection is pivotal.
- Define clear goals and purposes for data collection and processing that drive attempts to improve the quality of data and the efficacy of the data exploitation processes.
- Have clear purposes and state them before data collection takes place. This will optimize the amount of data collected and affect positively the data processes.
- Collect only data that are really needed. The amount of data collected represents an important element which can influence the quality of data and their processing, the definition of what such appropriate amount is can only come out from an accurate consideration of the organisation's needs.
- Therefore, a detailed a priori definition of data collection and processing goals and purpose will enable the design and implementation of the appropriate technical structure to process and store data before data are collected and grant an effective data management strategy.
- Moreover, avoiding the collection and processing of large, unjustified amounts of data is one of the easiest way to save resources and optimize organisational processes. It also, implicitly, addresses a key principle behind data protection legislation.
- Consider time. The preliminary definitions of needs and purposes should take into appropriate consideration not only the current data usage and exploitation strategies, but also all potential future uses and the needs of all different data users and data controllers to access the data set. Questions of collecting appropriate consent from data subjects are key here.
- Preserve the level of quality achieved. To maintain the level of quality achieved after the initial planning and data collection phase, it is advisable to empower data subjects and offer them a high level of transparency about the data processing, data collection purposes and any change in future data uses or policies. Moreover, allowing data subjects to have access to and interact with their data will increase data subjects' level of confidence and trust, increasing the level of data quality and trustworthiness.
- A high level of transparency will also affect the overall data accuracy by allowing data subjects to be in control of their data. If data subjects have access to their data, they can spontaneously collaborate with the organisation in keeping the data updated and accurate. Creating mechanisms and processes which remind data subjects where their data are stored and informing them about any change in data processing or data usage, will induce them to interact with their data and update them in case any major change occurs.
- Team work is key as data are generally handled by a number of persons within the organisation, each with their own different needs. It is essential therefore to focus on

organisational policies and put a strong emphasis on the importance of the data and the essential role that each and every individual plays in maintaining the quality of data. A sound information system is not enough if not adequately supported by a cultural change within the organisation. This supports the concerns raised in 4.7 about the culture within the public sector not being focused on making data available.

Specific consultation questions

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

13) Some key definitions, critical for the application of the PSI regulations are missing (such as what constitutes "public task" or what is an "interoperable licence"), whereas some definitions are problematic (e.g. "open data") and others are missing (e.g. data points) or it is not clear from which perspective they are given (e.g. is "information" defined from a legal or scientific point of view? How does it relate to data or documents?). In general, there is a lack of consistency in the vocabulary used by public sector bodies both within the UK, in the EU and worldwide with regards to issues of Open Data. There needs to be, at least in the context of the UK, an effort to produce a consistent vocabulary and taxonomy of terms related to Open Data consistent to (a) the regulatory environment; (b) the scientific and technological developments; and (c) the worldwide Open Data community. It is also important to clearly set the viewpoint from which the vocabulary is used and to resolve terminological differences found in different government discourses and documents (e.g. differences in the way "information" or "data" are used in the context of Open Government, PSI, FOI, Data Protection or IPR related policy documents).

14) The term "open data" in relation to public services is only defined with regards to the UK OGL. This substantially limits its scope and ignores the reality of other OGLs (such as the French OGL that is explicitly made compatible with all other OGLs the CC BY 2.0 and the Open Data Commons Licence) or the use of other standard licences, such as Creative Commons Attribution or the Open Data Commons Attribution licences. The reference to the CC BY licence made in the footnote of the definition, as it stands, only adds to the ambiguity, since it is not clearly stated whether it constitutes an Open Data licence in accordance to the definition.

15) It is highly recommended that the definitions section becomes more comprehensive and consistent. Also, the question of what constitutes Open Data should become a separate section, aiming specifically at licensing and defining which licences are Open Data licences and which not.

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

16) All third party IPRs have to be cleared.

17) Consent has been obtained from the data subject (ideally) or data have been anonymised (but see concerns above).

18) Data should not fall under any of the restricted categories found in the PSI Directive.

19) All other data–sets have to be made open.

20) Priority should be given to data sets that are crucial for crisis management and prevention.

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?

21) The overall aim of releasing data should be not to gain from fees but from the overall economic activity produced as a result of the data release and the reduction of transaction costs for data– re–users

22) The requestor should only be asked to pay for data that are produced outside the normal activities of the public body and there the costs should be kept as close to marginal costs as possible.

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?

23) The criterion should relate to the funding of the relevant public service providers: to the extent that the production of the data is funded with public data, such data have to be made open.

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

24) There is need to associate the release of open data and related tools to the specific public service providers and reward them accordingly either with budget allocation scheme or in other ways through a data citation indexing mechanism: the more data they release, in better quality and the more they are reused, the more they should be rewarded.

25) There is also clearly a relationship between transparency issues and its impact on the trust in open data by data subjects.

An enhanced right to data: how do we establish stronger rights for individuals, businesses and other actors to obtain, use and re-use data from public service providers?

26) There is need to establish a positive to access and reuse PSI.

27) The right could be shaped in the form of an enhanced FOI right. However, there are important questions of how such a right would be enforced. Thus, more information about the right to challenge decisions not to publish (would this independent body be the ICO and if so, would the commissioner be given stronger powers to insist on

publication?). More generally there is a question of how whether a 'requirement' would be sufficient to change culture such that proactive publication of data became the norm.

28) It is necessary to clearly define the scope and ambit of the right, i.e.:

a) What the subject matter will cover (will it equal current PSI or go beyond that?). It is strongly suggested that cultural sector and scientific information is also included.

b) What the actual rights of the beneficiaries will be.

29) A system of enforcement and implementation has to be clearly set in process. It is suggested that a separate independent regulatory authority is set up to administer and enforce the new right to open data is established. This authority might sit within the existing Information Commissioner's Office as it would clearly need to work closely with their role on FOIA.

Policy challenge questions

30) The question of ensuring procurement rules include consideration of making data available is important but is probably at a different level of detail to the rest of the consultation. Similarly, there is ambiguity between a requirement to procure systems that make "open by default" and the suggestion that this should be a "most attractive option for procurement".

31) One way of addressing this issue might be through the issue of model clauses for procurement contracts.

Setting Open Data standards

32) Paragraph 8.8 of the consultation is, in some ways, emblematic of the cultural problems faced in moving to an open data agenda as it states, implicitly, that some government processes are currently collecting and operating on data of such a low quality that it would be unhelpful to make this available to the public (yet it seems sufficient for internal use).

33) Recognising that, in practice, there is very little 'raw' data that is likely to be disclosed (for example, data is frequently 'normalised', error checked etc.) it would be helpful to include details of any such normalisation processes used as well as the 'raw' data themselves, when they are made available.

Setting transparency standards: what would standards that support an enhanced right to data among public service providers look like?

34) The standards should cover:

Legal level:

a) standard licensing schemes;

b) basic open data principles;

c) data accreditation schemes on the basis of the open data principles;

d) open licence framework;

e) standard privacy policy notices for public bodies;

f) organisational operational procedures;

g) clearance of IPR;

h) personal data compliance;

i) obtaining and managing consent (ideally) or establishing effective anonymisation procedures.

Technical level:

j) use of persistent identifiers;

k) use of open formats and standards ;

l) compliance with open data standards set by international bodies.

Corporate and personal responsibility: how would public service providers be held to account for delivering Open Data through a clear governance and leadership framework at political, organisational and individual level?

35) Each public body should release a data management strategy setting specific goals for opening up data and publicly announcing the criteria according to which they will assess themselves regarding the success of their open data schemes.

36) In order to increase accountability, each data set has to be assigned to specific custodians which should be personally accountable for the status and quality of the specific data set

37) There needs to be a census on the existing data sets and their quality and all data and the data custodians should be indexed.

38) With regards to the processing of personal data, there should be proactive measures to increase transparency in the way they are processed. Public bodies in their Privacy Impact Assessments should also include risk assessments regarding the possibility of violating personal data protection rules through data mashing and accordingly make informed and transparent decisions as to which data sets to release and how they are to be released.

Meaningful Open Data: how should we ensure collection and publication of the most useful data, through an approach that enables public service providers to understand the value of the data they hold and helps the public at large know what data is collected?

39) There needs to be more comprehensive monitoring of the use of the data and qualitative as well as quantitative studies regarding their use and re–use. An obvious relationship exists between the use and re–use of the data and its underlying quality. Other forms of monitoring could include:

a) Statistics regarding the performance of different public bodies in releasing data sets;

b) Quality assessments based on clear and transparent criteria as to what data quality amounts;

c) Pro–active engagement of the market and the open data community in order to obtain feedback as to how data are re–used;

d) Creation of case–study libraries with regards to the re–use of open data that could be completed by businesses and individuals;

e) Creation of open data indices containing evaluation of open data performance of different public bodies;

f) Establishment of capability maturity models for public sector organisation in order to openly release data. It is important to note that while there are evaluation schemes such as the open knowledge index that assess open access capacity in the national level, there is nothing equivalent for open data in the national level and certainly no scheme to assess capability and assist maturity in the organisational level. Such instruments are essential to support public bodies in their efforts to release open data;

Government sets the example: in what ways could we make the internal workings of government and the public sector as open as possible?

40) Have clear, transparent and coherent policies for the processing and dissemination of data sets.

41) Provide self assessment schemes.

42) Have transparent pricing schemes for all commercial information.

43) Provide information regarding the re–use of data sets.

44) Provide information regarding procurement decisions, specifically with regards to software tools for the re-use of PSI.

Innovation with Open Data: to what extent is there a role for government to stimulate enterprise and market making in the use of Open Data?

45) Use crowdsourcing models for identifying important data sets and supporting new ideas for the use of the open data.

46) The role of the government should be to provide the data sets and support the individuals and organisations wishing to re–use them but not to engage itself in the offering of value added services beyond their public task obligations.

Dr Edgar A. Whitley

On behalf of the EnCoRe project.