Adobe Systems Incorporated
345 Park Avenue
San Jose
CA 95110-2704
USA

Adobe Systems Incorporated response to HM Government Consultation

"Making Open Data Real: A Public Consultation"

Introduction

Adobe Systems Incorporated welcomes the opportunity to respond to this public consultation, and to contribute to the best possible implementation of the UK's Open Data policy. Given the UK's leadership role on Open Data, a balanced and effective policy will be essential in driving best practices not just in the UK but across Europe.

Adobe is responding to this consultation in its capacity as a provider of potentially relevant software technologies: as the largest supplier of PDF technologies, we have a stake in ensuring that the full capabilities of that format are well understood by the Open Data community; on the other hand, many of the applications and online services developed using open data are likely to be created using Adobe's suite of platform- and device-agnostic developer tools.

The Consultation document highlights very effectively the overlap between Transparency and Open Data. Indeed the ease with which concepts like "data" and "information", or "citizen" and "developer" are used interchangeably highlights the very broad objectives of the policy. It is therefore striking that the consultation so readily commits to promoting "machine-readable" outputs over human-readable outputs, suggesting a preference for providing "data" to "developers" over the provision of "information" to citizens". We believe that these concepts should not be in conflict, and that going too far in one direction to the detriment of the other could be potentially wasteful for government. Raw data, on its own, is of limited use to a very limited set of stakeholders. Care needs to be taken to provide meaningful context so that the respective data can be interpreted and re-used in an appropriate way.

In our answers to the questions below we aim to show how PDF can actually be an effective bridge between the twin demands of transparency and reusability, providing a handy format for the publication of both machine-readable data and human-readable information. It is well known that, with the proper tools, the contents of a PDF file can be extracted into different forms, including XML. But what is not so well known is that PDF files may contain attachments, such as done with e-mail. For example, a PDF file could contain both the raw XML data, the XML schema to understand the data, and some additional information to give that data context as well as a human-oriented presentation of the data. Standard PDF software products have simple user interfaces for attaching and extracting data files from a PDF "container."

Far from being part of the problem and "locking data in", PDF can help governments to leverage existing IT investments, and ultimately reduce barriers to publication and re-use of public sector data. We urge any recommendations on the technical aspects of an Open Data policy to remain technology-neutral, and allow existing formats to contribute to the development of a culture of Open Data in the UK.

Response to selected consultation questions

Questions for Consultation *(Page 7)*

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

While the definition of Open Data outlined in the consultation document Glossary provides a good basis for an open data policy, we believe further thought needs to be given to the draft Open Data principles reproduced in Annex 2.

Clearly, in an open data context the requirement that data should be made available in "reusable, machine readable form" is not controversial. However, giving machine-readable outputs priority over human-readable outputs, rather than establishing equality between the two concepts, may not be the optimal outcome for government. Raw data is useful to a very limited group of people, perhaps only the originators of the data. Ensuring that data / information is available in ways that are both machine-readable and human-readable is likely to satisfy a wider group of citizens and contribute simultaneously to policy goals relating to transparency and information reuse.

Furthermore, the apparent exclusion of PDF as a suitable format for use in this context appears gratuitous, and is grounded in an incomplete understanding of the capabilities of the format. As we explain elsewhere in this consultation response, PDF is in fact both a human-readable AND a machine-readable format. The contention that "government information is locked into PDFs" shows that this is not understood. It also clearly pre-empts the outcome of any properly consulted and well-researched analysis of appropriate formats for open government data reuse. As such it risks marginalising one of the most available, non-proprietary, open formats available to government entities, with the potential to contribute a great deal to the promotion if Open Data in the UK.

Similarly, while the recommendation that "public data will be published using open standards" is not controversial (PDF is an open standard, published by ISO - ISO 32000), the presumption in favour of "the relevant recommendations of the World Wide Web Consortium clearly pre-empts any wider discussion about appropriate formats and their use cases for data publication. While we appreciate the need to develop a policy that achieves real results, we respectfully urge the UK government to explore further the implications of standardizing too soon around any given format in an area of high innovation like Open Data.

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

Publication of data should be as easy as possible for public service providers. Anything which increases the friction of data publication is likely to have the effect of inhibiting publication. Among the technical considerations, the question of allowable formats for the publication of government data PDF is often raised as an issue. PDF, an open standard owned by ISO, is capable of supporting both machine-readable and human-readable use scenarios and is widely installed across UK government. Its capabilities – explained in more detail elsewhere in this consultation response - are not widely understood. Any guidance or mandates on allowable formats for the publication of government data which tend to prevent the use of PDF are therefore likely to act as an inhibitor to the publication of open data: from an economic perspective, technical mandates could prevent government bodies from leveraging existing IT investments; from a human capital perspective, they would fail to harness existing skills among public sector employees.

An enhanced right to data

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?

Any right to data should allow government bodies to leverage existing IT and skills investments as far as possible. A right to data which avoids re-using existing IT infrastructure and building on the existing IT skills of the public sector workforce is less likely to achieve quick, widespread adoption than a solution which requires the adoption of new systems and requires additional skills and training. In developing an open data policy, and a right to data, reuse of existing IT resources and human capital is a key factor in reducing procedural burdens and managing costs.

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

It is not clear exactly which "Open Data standards" are referred to in this context, although the consultation document does refer to the work of the Public Sector Transparency Board (PSTB) in promoting Open Data standards. By extension we assume that reference is being made, inter alia, to the PSTB's "Draft Public Data Principles".

The PSTB principles make an explicit – and unexplained - recommendation against the use of the PDF format in the context of open data. We believe that this recommendation is grounded in an incomplete understanding of the capabilities of the format.

PDF is an Open Standard, as published by ISO (ISO 32000). There is an active community of PDF tool developers – around 1800 – offering widely-available, commercial and open source tools, many of whom are actively working on ways to make PDF even more useful in an open government context.

As we explain in more detail below, PDF is in fact both a human-readable AND a machine-readable format. The ability to extract data is included in the ISO3200 specification. We therefore believe that the recommendation against PDF unnecessarily pre-empts the outcome of any properly consulted and well-researched analysis of appropriate formats for open government data reuse, and would welcome the opportunity to discuss this matter further with government. The draft recommendation, as it stands, risks marginalising one of the most widely-available, non-proprietary, open formats available to government entities, and one which therefore has a great deal to contribute to promoting Open Data in the UK.

This example highlights the risks of establishing specific guidance within procurement processes in relation to a specific policy outcome such as Transparency or Open Data. Given the misconceptions surrounding the capabilities of PDF cited above it seems likely that allusion is being made to the format when, on page 25, the consultation document refers to "data…stored in a fashion that makes it difficult to extract". We therefore urge UK Government to better understand the full capabilities of the PDF format before making any pronouncements about formats in a procurement context.

The following examples demonstrate why PDF is actually one of the main tools at the disposal of the UK government for making "data extraction easier and cheaper."

Examples of PDF capabilities in an Open Data context:

- **PDF as an all-purpose envelope**
  Raw data needs a context or metadata (information about the data) to be of use. Any number of file attachments, in any format, can be embedded into any PDF file and extracted for use by anyone receiving the PDF file. In addition, when files are attached/embedded into the PDF, they will be compressed using the

same compression method that is used in ZIP files. For the purposes of distributing government data, this is nearly ideal. The PDF file can carry raw data files (in xml or any format) as compressed attachments and all the additional semantic information that would be needed in order to make use of the data — the metadata. If the raw data is in XML form, then a compressed XML Schema file (.xsd) can also be attached to the PDF document. This sample PDF envelope starting from this government dataset demonstrates this functionality. Note that both the XML dataset and the associated Schema file are attachments to the PDF that helps to define the XML markup language used for this file. We took the general introduction from the government web page and made up a brief description for each of the XML elements found in the file.

- **Programmatic extraction of data from any PDF**
  PDF attachments can be of any format and can also be organized hierarchically, just as you can with a ZIP file. There are standard products that provide simple user interfaces for the creation, modification and viewing of PDF including insertion and extraction of PDF attachments. And like ZIP, there are numerous open source projects devoted to the processing of PDF. One very popular one is iText by Bruno Lowagie and iText Software. They also have an excellent example demonstrating how to use iText to create a PDF containing various data files. In addition, it also shows how someone could **programmatically** extract those contents. This means that the PDF is not only usable for pure human consumption but also for pure machine consumption or a combination thereof. Because PDF is an ISO standard, anyone is free to develop such solutions at any time around the PDF format. We believe that this proof of concept comprehensively addresses the main criticism of PDF among the development community that data is "difficult to extract". The opposite is in fact true. See this post from the InsidePDF blog for more context.

- **Enabling hybrid online / offline solutions**
  The consultation rightly refers to linked data, and storing data online. This example from Eurostat[1] shows how PDF can be used to promote both transparency (by presenting data in a meaningful context for citizens) and data reusability (by making the raw data available for developers). The PDF doc gives hyperlinked access from a blue link underneath the tables to a website containing the raw data behind the graphs for those that need to interrogate the data further. Again, data is not trapped into PDF, and those looking for human and machine-readable data can be satisfied.

For these reasons, we do not necessarily share the view that there is a choice to be made between publishing data on the internet or storing it in local files or on paper. Equally, the assertion that embedding open data in the UK will require "replacing outdated data management systems…(with) fundamentally new tools that end the classic model of saving files to network drives" appears to us to be somewhat overstating the need for change. In fact, as PDF has shown, the tools for achieving open data (whether for machine or human consumption) *are already available and in the hands of government employees across the UK.* All that is needed is that policymakers and government officials understand better their capabilities, and start to take advantage of them. Our approach is therefore more one of evolution than revolution.

Setting transparency standards

Q1. What is the best way to achieve compliance on high and common standards to allow usability and interoperability?

As the consultation rightly points out, Open Data is a means while the Objective is transparency. However, Open Data without information is not meaningful or useful. Contextual information may therefore be as important as the raw data itself in achieving transparency.

[1] http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-EI-11-001/EN/KS-EI-11-001-EN.PDF

Open Data is a concept in its infancy. We therefore believe that the most efficient way of ensuring compliance with new Open Data requirements is to approach the issue from the perspective of use cases. While there may be specific circumstances where it is possible and appropriate to mandate the use of specific formats for specific use cases, this is not likely to be the case across the board. The consultation appears to recognize this when it talks of "recommended publication formats *appropriate to the context*" (emphasis added).

On the other hand the consultation document clearly favours, by implication, adopting a single open data standard. A single standard, whether from the W3C or any other body, is unlikely to provide the flexibility needed to take into account all the different use cases or the many different contexts in which the data is produced and made available. Standardizing too soon around a specific technology solution may not, therefore, be wise, especially if this also prevents public sector entities from leveraging existing investments, or forecloses the ability of existing formats to contribute to the open data revolution.

We believe that this question is right to focus on the term "usability". We believe that this term should take into consideration the needs of all citizens, not just developers using raw data. Usability in this broader sense means that there is a need to provide adequate context to the data being published. Otherwise the usability being promoted benefits only a narrow community. However, the description of the levels within the indicative Five Star Rating for Open Data suggest that in fact only a very narrow concept of usability is being addressed. Nowhere is reference made to the context or presentation of the data, or its potential use by citizens at large. The example given above of PDF as an all-purpose envelope, on the other hand, demonstrates how the two concepts do not need to be in opposition. For example, a PDF file could contain both the raw XML data, the XML schema to understand the data, and some additional information to give that data context to make the 3* level readable by humans as well.

Corporate and personal responsibility

1. How would we ensure that public service providers in their day to day decision-making honour a commitment to Open Data, while respecting privacy and security considerations.

While we believe that there is a role for the Public Sector Transparency Board, we would welcome moves to ensure broader consultation with interested stakeholders on a regular basis. The Draft Open Data principles, while largely sound, significantly misrepresent the capabilities, and suitability of, the PDF format in the open data context. Such inaccuracies are easily avoided by broadening the channels of consultation.

Meaningful Open Data

5. Should the data that government releases always be of high quality? How do we define quality? To what extent should public service providers "polish" the data they publish, if at all?

Depending on the context, there may be a number of additional considerations before public sector service providers publish data:

- **Guaranteeing authenticity –** users need to trust that the data is from the source they believe they are engaging with. They may also need to know that the data is from a particular date. An additional advantage of using PDF envelopes is that the digital signature technology available for PDF files can cover the attachments automatically since they are an official part of the PDF file. Government agencies can send digitally certified PDF files containing data files and their customers can authenticate that the PDF, *and all the attachments*, came from that agency and have not been tampered with.
- **Managing risk –** it is impossible to guarantee that a Right to Data or an automatic presumption that data will be made public will not have unintended consequences. Automatic publication of data without additional

context could, in some circumstances, give rise to complaints or criticism (particularly on sensitive matters like public expenditure). Adding context to data may therefore be desirable in some circumstances. An Open Data policy should not have the effect of *preventing* the addition of contextual information alongside raw data which might clarify the significance of public data.

Innovation with Open Data

1. Is there a role for government to stimulate innovation in the use of Open Data? If so, what is the best way to achieve this?

There undoubtedly are many areas where government has a leadership role to play in stimulating enterprise and creating a market in the use of Open Data. We urge extreme caution, however, when government steps into the area of technical guidance on formats and standards. For the reasons we have already stated, IT procurement mandates based on incomplete information risk harming the marketplace. Efforts to standardize too soon on technical standards in an area of high innovation and fast-paced development also risk having an outcome that is more negative than positive. We urge technological neutrality, and respectfully remind government of the need to consider re-use of information and transparency as two complementary, not opposing, principles of equal weight.

Conclusion

Adobe welcomes the exciting developments in the Open Data movement. As a major technology provider we understand that we have a stake in working with all stakeholders to provide the solutions they need. This means listening to the developer community to ensure that extracting data from our PDF formats is as frictionless as possible; or providing creative tools that help developers build applications or websites once and roll them out on different devices with minimum difficulty; and it means standing by our existing government customers and helping them reuse their existing investments and skills to ride the Open Data wave while ensuring maximum value for money. When the interests of all sides are balanced then we have every faith that the "6 opportunities" outlined in the consultation are all there for the taking, and that the UK can be a global leader in Open Data.

** *end* **

For more information contact:
John Jolliffe
Senior Manager Government Relations
jjolliff AT adobe DOT com