

Making Open Data Real: A response from the DPC

About us

1. The Digital Preservation Coalition (DPC) is an advocate and catalyst for digital preservation, ensuring our members can deliver resilient long-term access to digital content and services. We are a not-for-profit membership organisation and our primary objective is to raise awareness of the importance of the preservation of digital material and the attendant strategic, cultural and technological issues. We support our members through knowledge exchange, capacity building, assurance, advocacy and partnership. We draw our members from the UK and Ireland and have series of alliances with organisations across the world. Our vision is to make **our digital memory accessible tomorrow**.

Summary of our response

2. The DPC believes that open data should be planned for the long-term otherwise the opportunities that it creates will be unsustainable and underdeveloped. Four practical implications follow from this principle: open data needs to be signposted predictably so that links and references to data are resilient; open data needs to be robust in terms of format, media and description to avoid the inadvertent disruptions caused by obsolescence and media failure; changes to open data need to be tracked and published to ensure that the integrity and authenticity is not lost; open data needs to be predictable in form enabling comparison of performance through time and facilitating the creation and refinement of analytical tools.

3. In short, we believe there is **no meaningful access without preservation**. By preservation we mean all of the actions required to maintain access to digital materials beyond the limits of media failure or technological change. Those materials may be records created during the day-to-day business of an organisation; "born-digital" materials created for a specific purpose; or the products of digitisation projects.

4. We **welcome the explicit statement about the value of data**. We note and support the focus on access and the benefit and value that flows from this. We seek **clarification on the scope** of data and agencies likely to be drawn into this proposed legislation. **We offer our assistance in identifying and resolving issues of resilience and continuity that may arise**.

An enhanced right to data

5. You asked, 'How do we establish stronger rights for individuals, businesses and other actors to obtain, use and re-use data from public service providers?'

6. We think that the presumption in favour of publication could be strengthened by explicitly requiring that publication plans include a reasoned and viable statement about retention and deletion. It is our view that departments and agencies are right to dispose of data sets and should not be required to retain data beyond what is reasonable. However decisions about deletion and retention ought to be managed and not be simply left to chance or technological obsolescence. This might be done by introducing an analysis of data in terms of its immediate value (<1 year), short-term value (1-5 years), medium-term value (5-20 years) and long-term value (>20 years).

7. Assuming that assumptions about deletion and retention are strategically managed and made explicit, there is a role for an agency to endorse retention plans and ensure compliance with them.

8. We have no view as to whether the stronger presumption on publication includes sufficient safeguards to protect privacy.

9. The resource implications are likely to be specific to individual data sets depending on scale, complexity, demand and the quality of descriptive information around them. However previous experience tends to suggest that most of the actions required to provide long-term access – metadata creation, file validation, quality assurance, storage, data management, user monitoring and so forth – are tasks that need to be performed to provide meaningful access over the short term also. Other tasks such as preservation planning and technology watch are best done collaboratively and bring economies of scale. However, evidence shows that there is a distinct skills gap in the planning and delivery of long-term access (see Sharpe

and Waller 2006, 31, Boyle et al 2008, 5, Angevaere 2009, 17, van der Hoeven 2009, 16). This is likely to be the largest resourcing issue in the first instance.

10. The standards necessary to deliver open data for the long-term need to be embedded into the procurement of new ICT contracts. This may require appropriate skilling of those involved in procurement. Moreover, in developmental areas like digital preservation standards continue to emerge. Consequently procurement needs to be informed by a standards observatory and gap analysis functions.

Setting Open Data Standards

11. You asked, 'What would standards that support an enhanced right to data among public service providers look like?'

12. There are a number of means to achieve compliance with standards and the way these are implemented may need to be adapted to suit local environments. However, experience in digital preservation shows that early intervention is required in all but the most exceptional cases. In analogue technologies paper could be transferred to a physical repository where long-term access could be assured through a combination of preventative and active interventions: preservation was not an issue until a document had already been used. Digital technologies have much shorter lifecycles so the need to act is considerably more urgent and the fundamental data required to ensure preservation becomes more difficult to gather through time. Digital archaeology – the process of reconstructing inaccessible data from disintegrated traces - is highly specialist and unlikely to be affordable or appropriate in all but the most unusual cases. Therefore, open data needs longevity embedded at the point of creation and planned at some point before that.

13. DPC has no particular view on how to collect data about the experience of users or the accreditation of information intermediaries.

14. It is a cause for concern that the standards listed, such as the '5 Star rating for Open Data' have no temporal dimension. Access needs to be managed through time otherwise links will break; descriptive information about how to use the data will become lost; formats and media become obsolete making access impossible or impractical; authenticity compromised; and time series disintegrate. This is a challenge to the vision for open data which can be easily resolved but which must be addressed as a matter of urgency.

Corporate and Personal Responsibility

15. You asked, 'How would public service providers be held account for delivering open data through a clear governance and leadership framework at political, organisational and individual level?'

16. We have no specific view on how these requirements should be met, or about how privacy and security considerations should be implemented in the day-to-day decision making of organisations.

17. Assuming that personal responsibility for the right to data resides at board level within organisations, then it is important that board officers understand the long-term implications of their policies. They should ensure that the retention and deletion are consistent with their strategic goals and that policies are implemented.

18. We have no view on sanctions or on the perceived benefits of a Sector Transparency Board. However we believe that if it were to be established, such a board and the minister who over sees it should be properly equipped to ensure that the right to data is an enduring one.

Meaningful Open Data

19. You asked, '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?'

20. We have no specific view on the use of data inventories or what should be included in them and what data government should collect and publish.

21. The value of data is likely to change through time and that temporal comparison is likely to be a particular value which would be curtailed if consideration is not given to the potential of data through time. Consequently the value of data needs to be considered over an immediate term (< 1year), short term (0-5 years), medium term (5-20 years) and long-term (more than 20 years). Such a planning horizon will ensure that the long-term exploitation of value is considered by those responsible for it.

22. We have no view on what data the government collects unnecessarily. It is our view that data can be disposed of once it has reached the end of its useful life. Consequently if the government seeks to identify 'unnecessary' data then it would be helpful to consider immediate, short, medium and long-term value.

Government sets the example

23. You asked, 'In what ways could we make the internal workings of government and the public sector as open as possible?'

24. We have no view on whether data is best accessed through a central portal or through a series of departmental portals, though our experience suggests that there is merit in distinguishing data and preservation functions from access and presentation functions. Such a distinction would encourage providers to develop their own interfaces to the data. The question of department or central portals therefore seems redundant. However we would strongly recommend that identifiers for data sets are persistent so that links are not inadvertently broken when data sets move.

25. We have no view on the prioritisation of data sets for publication or on whether a small and narrow or broad and shallow set should be the early priority. However, the fact that a dataset has been prioritised as being of specific value only serves to underline our basic contention about the need for long-term access.

Innovation with Open Data

26. We have no view on the role for government in stimulating innovation in the use of Open Data.

Points of clarification

27. We note that the scope of the data envisaged in the consultation is not clear. This creates the potential for misinterpretation and inadvertent complications. For example, the definition of data sets – described as structured and unstructured – could be interpreted to embrace a range of public records which are managed under legislation such as the Public Records Act, Environmental Impact Regulations and Freedom of Information. It is not clear whether the definition includes research data gathered and generated by universities and research councils in pursuit of publicly funded research: release of such data sets is governed by an intricate series of institutional, disciplinary and international contracts, mandates and policies. It is not clear whether the definition is intended to include library or archive holdings which may have been transferred to the public domain with expectations pertaining to embargoes or intellectual property rights. Therefore the scope of the data covered by the proposed legislation needs to be clarified. The wider the scope, the greater will be the preservation challenges and the more acute will be the capacity to deliver data on a resilient basis.

28. This lack of clarity on the scope of data creates further uncertainty about the agencies likely to be drawn into the legislation. Public data is managed by and created a broad range of suppliers who use it to configure, maintain and monitor a wide range of public services. Access to such data is implied by the government's broad policy objective but the extent of that reach is not clear. It is not clear if data gathered and managed by local government is in scope; it is not clear if private contracting agencies that provide public services are within scope. If so, the proposal implies a significant change in regulatory requirements for a wide range of public and private agencies. We have no comment as to where these boundaries ought to be set, but the farther the reach of the legislation, the greater the preservation challenges and the more acute the lack of capacity to deliver such requirements on a resilient basis.

Notes about this response

29. Although full members of the Coalition, The National Archives have not contributed to this document.

References

Angevaare, I, 2009 A Future for Our Digital Memory: Permanent Access to Information in the Netherlands (Interim Report, English Summary), NCDD, The Hague online at:

<http://www.ncdd.nl/en/documents/Englishsummary.pdf> last access 22/10/09

Boyle, F, Eveleigh, A and Needham, H 2008 Report on the Survey Regarding Digital Preservation in Local Authority Archive Services, Digital Preservation Coalition, York, online at

<http://www.dpconline.org/docs/reports/digpressurvey08.pdf> last access 28/08/09

van der Hoeven, J 2009 First Insights into Digital Preservation of Research Output in Europe, PARSE.Insight, online at: http://www.parse-insight.eu/downloads/PARSE-Insight_D3-5_InterimInsightReport_final.pdf last access 23/10/09

Waller, M and Sharpe, R 2006 Mind the Gap: assessing digital preservation needs in the UK, Digital

Preservation Coalition, York, online at: <http://www.dpconline.org/graphics/reports/mindthegap.html> last access 28/08/09

**Dr William Kilbride FSA,
Executive Director of the Digital Preservation Coalition
November 2011**