DigitalPreservationCoalition

Proposal on the Collection and Preservation of UK Offline and Microform Publications and UK Online Publications: A Response from the Digital Preservation Coalition

Introduction

The Digital Preservation Coalition (DPC) is a not-for-profit membership organisation whose primary objective is to raise awareness of the importance of the preservation of digital material and the attendant strategic, cultural and technological issues. Its vision is to make **our digital memory accessible tomorrow.**

We note and welcome the progress which has been made by the Legal Deposit Advisory Panel on recommendations for collecting digital materials. We are eager that the momentum recently achieved is maintained so that continuing progress can be made.

In summary, we **warmly welcome the proposal for regulation-based harvesting** and call for **early implementation** of this proposal. We offer the a**ssistance of the DPC in capacity building** for staff and tools which this will necessitate.

There is a range of opinions within the DPC's membership regarding the access provisions within the Proposals. These will be reflected in their own individual submissions to the consultation. The position of the DPC itself, however, remains clear that **future access to the harvested materials at any level will be impossible without the safeguards that rigorous attention to preservation provides.**

The membership of the DPC includes all five of the UK's deposit libraries and Trinity College Library in Dublin. Our members include important stakeholders with interests in web archiving such as the National Archives, National Archives of Scotland and the Public Records Office of Northern Ireland; content creators such as BBC and the Publishers' Licensing Society; public sector funders and commissioners of web content such as MLA, Scottish Arts Council, RCUK and JISC; users of web content such as RLUK; and specialist services already involved in web archiving such as The Wellcome Library, University of London Computer Centre and the Digital Curation Centre.

These comments have benefitted from consultation with the newly founded 'Web Archive and Preservation Task Force' which operates under the auspices of the DPC. The Task Force exists to share best practice with its members. It identifies, examines and reviews current policy in web archiving and preservation. It provides a mutually supportive environment for continued policy development for members and a mechanism through which non-members can engage with web archiving policy. In this way the Task Force helps to ensure that our generation can carve an appropriate legacy from the complexity and volatility of the web.

We offer our support in delivering appropriate elements of the Proposals.

Questionnaire

1. You asked **'What are your views on the options considered for this content?'**

   In our view, Option 2 offers a measured, sensible and achievable route to progress web archiving and preservation. Experience has shown that Option 1 - Permission-based harvesting - is cumbersome to operate and is not workable. Oftentimes, creators and publishers of content are not able to provide the permissions that libraries seek, are not in a position to respond to the request, or simply do not understand the proposition being put to them. This creates a considerable administrative overhead for the library and the publisher and it slows the process considerably. Moreover the resulting collection is based purely on responses from willing and able web masters. Consequently there is little prospect of developing a coherent national collection.

   Option 3 raises different questions and does not address the core issue. For example, in the private sector collecting and access policies would not be governed by open consultation and thus access to a representative legacy collection could be compromised. A business model based where publishers buy archival services could be sustainable: but the implied preferences that would accrue to specific interests would have unfortunate consequences and create a partial and selective memory. In fact, such services are already available but are not widely used and do not constitute a viable alternative to a coherent national collection. Indeed, the existing services of the UK Web Archive to some extent already offer this service: even when the service is available at no cost, it is under-used. It is worth remembering that the core challenge of web archiving in the UK, as identified by UKWAC, is the difficulty of obtaining permissions. This would remain unresolved.

2. You asked **'Are there any other options that should have been considered? If so, what are they?'**

   This consultation pertains to the legal and regulatory framework and we do not believe there are other options.

   There are a number of options for the underlying technology to harvest and preserve web content and experience suggests that this is likely to continue evolving. Consequently there is a need to ensure that the legal principles remain sufficiently independent of the operational infrastructure to ensure that deposit libraries are not inadvertently prevented from renewing and updating their workflows. We believe the current proposals protect this principle.

   It may be tempting to seek options from collaboration with partners internationally, or to attempt wider reform. These sound attractive but are unlikely to be achievable in the short term. International partnerships are likely to require the same set of changes in the legal and regulatory framework as proposed here.

3. You asked **'Do you agree with the analysis of these options? Explain why'**

Yes, we agree with the analysis of the options. We have given our reasons in response to question one.

4. You asked **'Do you agree that harvesting provides the most efficient and timely solution for deposit of publications in this category?'**
Yes, we agree that harvesting provides a timely and efficient solution to the deposit of publications of this kind. The scale of the task requires as much automation as is consistent with a well formed, well managed and accessible collection.

   Developments in the underlying technologies of web publishing and harvesting mean that the tools used to gather content are likely to change. But, provided the regulatory principles which entitle the deposit libraries to copy without having to seek permission is sufficiently independent of the technical infrastructure, the solution proposed here will be effective.

   We should remind the Panel that harvesting is only the initial part of the operation required to create a robust and trusted digital archive. Ingest procedures such as cataloguing, characterization, virus checking, quality assurance and storage will be required in the short term so that the library can obtain sufficient control of the collection. In the medium to long term additional services such as replication, refreshment and preservation planning will also be required and, depending on local practices, a degree of normalisation, migration or emulation may also be required.

5. You asked **'Do you agree that regulation is the most cost effective method of collection for the libraries and imposes no direct financial or administrative burden upon the publishers? Explain why.**
Yes, we agree that regulation is considerably simpler to implement for the libraries and imposes no direct financial or administrative burden on the publishers.

6. You asked **'Do you agree that this is an appropriate definition for the types of publications that should be included in the scope of regulations? Explain why. Is there anything that should be excluded from this definition?'**
Yes, we agree that this is an appropriate definition for the types of publication to be included in the scope of the regulations at this time. However we believe that a further effort will be required to address concerns about those publications currently out of scope and we call on the Legal Deposit Advisory Panel to address this additional area of work in a timely manner.

   The business case of charged internet services depends on valuable, desirable content. Consequently, the implication of the current proposal is that the most valuable or desirable content - collections from which posterity will likely benefit most - will be out of scope. It is reasonable that commercially valuable content be considered separately because of the implications this may have for legitimate exploitation and in any case the infrastructure for such a service will take time to mature. Nonetheless, we believe that commercial web content requires urgent attention. For this reason we urge the Legal Deposit Advisory Panel to sustain its momentum and promptly publish recommendations for the deposit of UK Commercial and Protected Online Publications.

7. You asked **'Do you agree with the territorial definition of the UK web? Explain why. Is there anything else that should be included in this definition? Is there anything that should be excluded from this definition?'**
   Yes, we agree with this territorial definition of the UK web. We believe that the definition offered here provides sufficient guidance for sensible decisions to be reached by curatorial staff.

8. You asked **'Do you agree with this analysis of UK web domain?'**
   We are not in a position to question or test the validity of the assertions made about the size of the UK domain, but believe that the logic used to establish it is sound.

   However the size of the domain causes us to note the scale of the operation that will be required in comparison to the current service operating as the UK Web Archives. Rapid development will require investment in staff and expertise as well as technology. See our answer to question 11 for more consideration of this topic.

9. You asked **'How do you see a Deposit Library driven system of web harvesting interfacing with a publisher driven duty to deposit under the 2003 Act?'**
   We see no difficulty in integrating the two models of a deposit library implied by the current proposal. There may be some benefits derived from de-duplication of content that is produced in both electronic and paper form, though in reality the cost of identifying these duplicates may be higher than the cost of retaining both and the decision on which version to dispose of may prove complicated. In any case the overlap is not great.

   However, the intention to develop proposals for archiving UK Commercial and Protected Online Publications may create a more complicated dialectic between the paper and electronic deposit. The prospect of being able to hold digital and paper versions of, for example, scholarly journals means that there will be a clear and more easily identified overlap between the two.

10. You asked **'How could deposit libraries most efficaciously ensure a comprehensive body of eligible content is deposited?'**
    This is a complicated issue which embeds three related areas of active discussion in the preservation community.
    The apparently simple solution is to effect a comprehensive harvest of the .uk domain. This would create a comprehensive and deep copy of the entire domain. But the size of such a harvest and the time required to process it would make it a slow option. Consequently, fast changing and culturally significant content would be gathered at the same interval as slow

changing or less significant pages. Ingest and data management procedures would quickly become stretched and the gap between harvests would be necessarily long.

A more subtle approach might be to identify key websites and to harvest selectively from the domain according to curatorial selecion. This ensures that the resulting collection matches the frequency of changes and is influenced by curatorial concerns over significance. However any human intervention is likely to be expensive in staff time and the selection process would require consultation and monitoring.

A combination of both approaches is possible.

It should also be noted that there are other agencies in the UK also involved in the archiving of websites and that collaboration with them will be mutually advantageous. For example the National Archives, the National Archives of Scotland and the Public Records Office of Northern Ireland have a range of statutory obligations to archive government web domain, while other interested parties such as the Wellcome Library, JISC, the BBC and the Parliamentary Archives have undertaken a variety of archiving of web based resources. Consequently actions should be congruent with the statutory and professional interests of cognate agencies.

11. You asked **'Do you agree with this costing model? Explain why. Are there costs that need to be factored in or excluded?'**
    Yes, we recognise the empirical base of the figures and the methodology used to derive them.

    We note that the profile through time seems to assume immediate and full scale implementation. While laudable we anticipate two barriers that will need to be overcome: the relative scarcity of skills; and the continued development of the technology. In addition we note the consequential risk and opportunity to the nation's digital preservation infrastructure

    Numerous recent surveys on preparedness for digital preservation (Angevaare 2009, Boyle et al 2008, van der Hoeven 2009, Waller and Sharpe 2006) show that skills are currently a key strategic gap in digital preservation. By implication it may be difficult to recruit appropriately qualified staff and it may be necessary to initiate an internal training programme for staff recruited to these roles. Because labour force development in digital preservation is a strategic requirement with wider relevance than the deposit libraries, there may be advantage in collaboration on this topic. Consequently, we offer our support in this area.

    Rapid development in web harvesting technology is to be expected during the implementation of these proposals. We have already noted the risk that the regulatory framework does not inadvertently prevent appropriate use of new tools and our belief that the proposals are successful in avoiding this risk. In addition, any subsequent implementation plan – including budget and timing – should include an appropriate resource to allow for the ongoing development of tools and their implementation.

Finally we note that the cost model for harvesting has implications for the wider costs base of electronic services within the deposit libraries which are included within these figures. In particular we are pleased that the digital preservation infrastructure has been factored into these costs. This will be required to ingest and guarantee continuing access to an extensive, expanding and heterogeneous collection. It is reasonable to suppose that the added expertise in web archiving will to some extent expand the capacity of the deposit libraries in digital preservation but the required investment in e-infrastructure for digital preservation is a strategic priority with wider relevance than web archiving and preservation. Developments in web archiving should contribute to and refine that strategic priority and should not deflect or postpone it.

12. You asked **'Do these assumptions adequately reflect the financial burden of publishers? Is there anything that needs included or excluded?'**
Yes, these assumptions adequately reflect the financial burdens of publishers, which is zero.

13. You asked **'Do you agree with the analysis of these options? Explain why?**
The consultation has a useful analysis of the options regarding the regulations about consultation within the premises of the deposit libraries. There is a range of opinions within the DPC's membership regarding the access provisions within the Proposal. Many members of the DPC hold deep concerns over the restrictions over access which are inherent in the recommendations. These will be reflected in their own individual submissions to the consultation. The position of the DPC itself, however, remains clear that future access to the harvested materials at any level will be impossible without the safeguards that rigorous attention to preservation provides.

14. You asked **'Do you agree with the analysis on making content available to the Deposit Libraries? Explain why. What else needs to be taken into account?**
Note our answer to question 13.

15. You asked **'Do you agree with this costing model? Explain why. What else needs to be taken into consideration?'**
Yes, we agree with this costing model, though see our commentary in question 11.

16. You asked **'Do you agree with this analysis of the costs and the impacts of each option? Explain why. What else needs to be taken into consideration?'**
Yes, we agree with this costing model, though see our commentary in question 11.

17. You asked **'Do you agree with the risks identified here? Explain why. Are there other risks that have not been considered? What would their impact be? Are some of these risks not really risks? Why?'**
    Yes, we agree with your analysis of the risks. We note some aspects of the risks may be mitigated by concerted and unified policy action across the UK web archiving community. Consequently the DPC Web Archiving and Preservation Task Force will provide timely and co-ordinated action to further refine responses to risk.

18. You asked **'Do you agree with LDAP's recommendation to regulate for this content? If not, what should be done instead?'**
    Yes, we agree with this recommendation for reasons given in our answer to question one.

19. You asked **'Do you agree with LDAP's proposed method for depositing content? If not, how else could this be done?'**
    Yes, we agree that this is the most efficacious way for depositing content. We note the need to distinguish between the tools used and regulatory framework in order that curatorial staff can be flexible in the light of emerging technology. We believe that these proposals protect that necessary distinction.

20. You asked **'Do you agree with LDAP's analysis of access provisions? Explain why? What other options are there?'**
    We repeat here our answer to Question 13. The consultation has a useful analysis of the options regarding the regulations about consultation within the premises of the deposit libraries. There is a range of opinions within the DPC's membership regarding the access provisions within the Proposal. Many members of the DPC hold deep concerns over the restrictions over access which are inherent in the recommendations. These will be reflected in their own individual submissions to the consultation. The position of the DPC itself, however, remains clear that future access to the harvested materials at any level will be impossible without the safeguards that rigorous attention to preservation provides.

21. You asked **'Do you agree with these cost assumptions? Explain why. What needs to be included or excluded?**
    We are not in a position to question the cost assumptions but we recognise their empirical basis and the methodology through which they have been derived.

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
March 2010