

Date: Monday 12th June 2006

Venue: [British Library Conference Centre](#)

The DPC held a one-day web archiving forum at the British Library. The first DPC web archiving forum was held in 2002 to promote the need to archive websites given their increasing importance as information and cultural resources.

Four years on, this event again brought together key web archiving initiatives and provided a chance to review progress in the field. The day provided an in-depth picture of the UK Web Archiving project as well as European initiatives. Technical solutions and legal issues were examined and the presentations encouraged much debate and discussion around different strategies and methodologies. The event made clear that the field has moved on tremendously from four years ago. The debate has broadened and so have the tools and methodologies.

The first presentation was from Philip Beresford, Web Archiving Project manager at the British Library [BL]. He spoke about the BL's involvement with UKWAC, the tools the project had built, the challenges they have had with the PANDAS software and the overall constraints of web archiving, especially as it is such a technology dependent discipline. Philip also outlined the web curator tool developed with the National Library of New Zealand and the next version of PANDAS. [UKWAC - the first two years \[PDF 33 KB\]](#)

Adrian Brown, Head of digital preservation at the National Archives followed on from Philip's talk as he outlined the future of UKWAC and its recent evaluation report. Adrian outlined the collection methods at the National Archives as well as database preservation and transactional archiving. He touched on one rather overlooked aspect, that of the long term preservation of the actual content. [Collecting and preserving web content \[PDF 401KB\]](#)

John Tuck spoke in the second session about the legal BL's deposit bill. He touched on issues regarding collection, capture, preservation and access to non-print collections. Of interest is how the legal deposit bill translates into the e-environment and web archiving; should web archiving extend to UK-related sites, not just UK-domain sites and are national boundaries less relevant now? He outlined the BL's two different strategies - taking a twice yearly snapshot of the entire UK web and the second being a more selective approach of sites that are deemed to be of national and cultural interest. He also stressed the lengthy permissions process that

gathering each web site entails. [Collecting, selecting and legal deposit \[PDF 42KB\]](#)

Andrew Charlesworth highlighted the complexity of the UK legal framework regarding web archiving. An emerging theme throughout the day was the debate about whether archives should ask for permission before or after they have collected websites. Andrew stressed the importance of understanding the regulatory framework. The field has moved on in that we know more today about the risks and benefits regarding web archiving than we did a few years ago. Any web archiving project probably needs to carry out risk analysis and to have insurance, in particular with regard the defamation law, ensure that they don't hold anything in their archive that could be used as legal evidence. [Archiving of internet resources: the legal issues revisited \[PDF 33KB\]](#)

Julien Masanes spoke about the European Web Archive [PDF 530KB] . He presented an interesting approach to web archiving - the information architecture of the web is such that its archiving should follow the natural structure of the web. Julien reminded the audience that web content is already digital and readily processable and that the web is cross-linkable and cross-compatible, a good foundation for an archive. He also stressed that web archiving requires functional collaboration. What is needed is a mutualisation of resources which combines competence and skills. [Internet preservation: current situation and perspectives for the future \[PDF 530KB\]](#)

Paul Bevan outlined the UKWAC project to archive the 2005 UK election sites. He described how three national libraries collaborated on this web archive. He touched on selection, collection remit for each library and frequency of snapshots. Did the general election classify as an event or as a known instance? Paul stressed the difficulties involved in obtaining permissions to archive electoral websites and the difficulties in identifying candidates websites. On a technical level the slowness of the gathering engine was also highlighted. [Archiving the 2005 UK General Election \[PDF 129KB\]](#)

Catherine Lupovici of the International Internet Preservation Consortium [IIPC] outlined the activities of the IIPC and outlined all the life-cycle tools that the team are working on such as ingest and access tools. She stressed the importance of collaboration in web archiving and it is clear that both UKWAC and IIPC do this successfully. [IIPC activity: standards and tools for domain scale archiving \[PDF 149KB\]](#)

The panel session was most productive. The panel leaders stressed that we are still in the early

days of web archiving. We can never be fully sure that the techniques employed are correct, but we have to make a start. However, more research needs to be carried out into the preservation techniques of the actual content. Access issues are also critical; searching a digital web archive won't employ the same search and retrieval tools as a traditional archive would and crucial access tools need to be developed for successful use of web archives.

On a technical note, we need to be aware of issues of browser compatibility in the future; there was a debate about whether it was an acceptable solution to obtain the source code of browsers in order to assist rendering pages in the future. It was highlighted that we have to be aware of unknown plugins which could hinder the readability of web pages. The importance of the ingest stage was stressed and the transformation of the digital object that should occur at this point to ensure readability. There may be legal issues to consider here however in transforming from one format to another.

Web archiving is not an isolated activity - so many web formats are now available as well as different content delivery mechanisms e.g. blogs and chat rooms. These formats make archiving even more challenging. There was a recognition that the community needs smarter tools to make web archiving scalable. There is definitely a need to semi automate quality assurance and selection. The question was raised whether or not we still need manual and selective archiving which is both time-consuming and costly compared to automatic sweeping of the web? The general consensus was that both methodologies should still be employed. The overall conclusion and recurring theme of the day is that collaboration is essential and no single organisation can carry out web archiving on its own. Projects such as UKWAC, IIPC and the European Web Archive demonstrate that much can be achieved in terms of solutions and methodologies.