date: Monday 25 March 2002
venue: IPMS, 75-9 York Road, London SE1 7AW

Web sites are an increasingly important part of each institution's digital assets and of this country's information and cultural heritage. This event, organised by the Digital Preservation Coalition (DPC), brought together key organisations in the field of web archiving in order to assess the needs of organisations to archive their own and others' web sites, to highlight good practice, and to influence the wider debate about digital preservation.

Meeting Report

This meeting report provides a short summary of the DPC Members Forum on web archiving held on 25th March 2002. Individual PowerPoint presentations from each of the speakers are available below

Web sites are an increasingly important part of each institution's digital assets and of this country's information and cultural heritage. As such, the question of their management and archiving is an issue which UK organisations need to be increasingly aware of. This event, organised by the newly-created Digital Preservation Coalition (DPC), brought together key organisations in the field of web archiving in order to assess the needs of organisations to archive their own and others' web sites, to highlight good practice, and to influence the wider debate about digital preservation.

Neil Beagrie, Programme Director for digital preservation at JISC and Secretary of the DPC, began the day's proceedings by welcoming delegates to the event, the first event on web archiving to be organised by the DPC. He stressed the importance of the issue.

The first speaker, **Catherine Redfern** from the Public Record Office (PRO) provided a short general introduction to web-archiving. Web sites are records, and as such, need to be managed and archived. However selection was necessary too, said Ms Redfern. But what are the criteria to be employed in such a process of selection? And how important is the capturing of the 'experience' of using the web site given that the look and feel of a site are an intrinsic part of the record. It was important, concluded Ms Redfern, to accept that perfect solutions do not exist, and that flexibility means that it may be the case that different solutions existed for different web

sites.

Brian Kelly of UKOLN followed and covered the size of the UK web domain and the nature of UK websites. He emphasised the sheer scale of the challenge by looking at definitions and measurements of the UK web space. A number of different approaches by organisations came up with different measurements, but a figure of 3 million public web servers which contained .uk within their URLs was given by Netcraft. In 2001 OCLC's Web Characterization Project suggested the UK accounted for 3% of the sites on the WWW. Searches using AltaVista further suggested that UK websites might contain around 25 million pages. Preserving web sites which we are unable to count will prove particularly difficult, he said, but perhaps the most important question was: at what rate is the UK web space growing?

Brian Kelly then went on to describe issues encountered during work on the UK webwatch and a pilot study to explore archiving issues for project websites funded under the JISC's eLib programme. He also described the Internet Archive (www.archive.org/) which is building historical collections of web content. He concluded that measuring the size of the UK web is difficult but the experiences of web harvesting robot developers and web indexers will provide valuable information for archiving the UK web.

Comparisons with other international situations are important in this context, and **Julien Masanes** from the Bibliotheque nationale de France (BnF), gave the French perspective on these questions. In France the Government is currently in the process of modifying the law regarding legal deposit of online content. Masanes explored the issue of archiving the "deep web" generated by databases and mechanisms for improving the current generation of web harvesters. The BnF is currently researching the best way to manage procedures of selection, transfer and preservation, which could be applied on a large scale within the framework of the proposed new law. Two large-scale projects are proposed as part of this ongoing research. The first one has begun and involves sites related to the presidential and parliamentary elections that will take place in Spring 2002 in France. More than 300 sites have already been selected and the BnF collects about 30 Gb per week. The second project will be a global harvesting of the '.fr' domain in June.

If the sheer scale of the amount to be archived presents a major challenge, it is one that the BBC, with a million pages on its web site, and each regularly being updated, faces as a matter of course. **Cathy Smith** New Media Archivist of the BBC spoke about modernising the BBC archive to include its digital content and the huge logistical and legal problems that this can involve. The BBC's Charter responsibilities mean that it must archive its content, while its

message boards, live chat forums, etc. mean that Data Protection becomes a serious issue in this context too. Multi-media content, often created through non-standard production processes, add further problems while proposals to extend the period within which the public can make formal complaints from one year to three years, has important consequences for the amount that will need to be archived. Ms Smith talked of the need to change perceptions from archiving to media management and for more pre-production emphasis on generating metadata and considering future re-use. She also emphasised the fact that the archive needs to recreate the look and feel of the original record since this was an important aspect of what it is that the BBC does.

A number of short reports from DPC members followed in the afternoon focussing on current initiatives and pilot projects. **Stephen Bury** of the British Library spoke of the BL Domain UK pilot project, the capture of 100 websites, and some of the criteria used by the BL in its current archiving activities, given the lack of legal deposit requirements. These criteria include topicality, and reflecting a representative cross-section of subject areas. Metrics of sites captured were provided, for example only 10% were "Bobby" compliant. Future developments would include scaling up the project and international and national collaborations.

Stephen Bailey, Electronic Records Manager for the Joint information Systems Committee (JISC) spoke of the JISC's efforts to implement its own recommendations in electronic records management and its current project of redesigning its own web site. The archive module of the new website will allow for identification and retention of key pages and documents and will also allow a greater degree of functionality for end users. Centralised control of the web records' lifecycle will allow for greater uniformity but will place demands on content providers. Future developments will include working in partnership on long-term preservation with the national archives and libraries and looking at preservation of the distributed JISC-funded project websites.

Steve Bordwell of the National Archives of Scotland asked whether we should even be attempting to preserve web sites in terms of look and feel, and whether we should rather be focussing on their content. He discussed their first work in the field, archiving snapshots of the Scottish Parliament website and the "Suckler Cow Premium scheme", a website based on an Oracle database with active server pages. They cannot preserve the whole application for the Suckler Cow site but will capture and preserve the dataset and use screencams to preserve the look and feel.

David Ryan of the PRO looked at the project to preserve the No. 10 web site from election day June 2001, and asked what an acceptable level of capture and functionality might be in terms of

archiving and preservation procedures. **Kevin Ashley** of the University of London Computer Centre (ULCC), suggested that we need to think what the purpose of websites is precisely and what their significant properties are in order to formulate criteria for selection, capture, and preservation.

Robert Kiley spoke about the joint Wellcome Trust/JISC web archiving feasibility study and the specific part of this that is looking at archiving the medical Web. He emphasised what we are in danger of losing if no action is taken. Once again, the sheer volume of the medical Web presents significant problems for selection: quality would be one criterion, but how should we judge quality? In addition, many database are published only electronically, while discussion lists and e-mail correspondence are also potentially of immense importance to future generations of researchers. Are the next generation of Watson and Crick already communicating electronically via a public email forum and will this survive? He outlined key issues to be addressed by the consultancy including copyright, costs and the maintenance of any medical web archive.

Concluding discussion on the future way forward for the UK emphasised the value of sharing current approaches and technical developments on web archiving both internationally and within the UK. There are still many technical challenges including the preservation of database driven sites and the need for better tools for harvesting and archiving webpages. It was recognised that the scale of the task in the UK was significant and would require careful selection of sites as well as collaboration between organisations, to address it effectively. The DPC would be setting up further individual meetings between members to advance discussions initiated at the forum and to develop plans for scaling up current pilot activities.

End of Meeting Report

To open PDFs you will need Adobe Reader

Presentations

Session 1

Web-archiving: an introduction to the issues (PDF 17KB) Catherine Redfern PRO (based on MA research)

Developing a French web archive (PDF 115KB) Julien Masanes Bibliotheque Nationale de France

The UK domain and UK websites (PDF 292KB) Brian Kelly UKOLN

Archiving the BBC website (PDF 15KB) Cathy Smith BBC

Session 2

Members' contributions

Stephen Bury (British Library) (PDF 17KB)

Steve Bailey (Joint Information Systems Committee) (PDF 16KB)

David Ryan (Public Record Office) (PDF 201KB)

Kevin Ashley (University of London Computer Centre) (PDF 10KB)

Robert Kiley (Wellcome Trust Library) (PDF 127KB)

Steve Bordwell (National Archives of Scotland) (PDF 45KB)