Document Distribution Note
Release to Participants: Immediate
Release to Members: Immediate
Release to Public: 01/09/09

missing links : the enduring web

21/07/09 | The British Library Conference Centre, London

Conference Report

1. Introduction

The web runs at risk. Our generation has witnessed a revolution in human communications on a trajectory with the origins of the written word and language itself. Early web pages have an historical importance with prehistoric cave paintings or proto-historic pressed clay ciphers. They are just as fragile. The ease of creation, editing and revising gives content a flexible immediacy: ensuring that sources are up to date and, with appropriate concern for interoperability, content can be folded seamlessly into any number of presentation layers. How can we carve a legacy from such complexity and volatility?

Key issues for long-term access and preservation remain unresolved. How can content creators make sure their creations are durable without impairing their flexibility? How does web-archiving relate to data-curation and traditional archiving? What constitutes an appropriate legacy from a web site? What audiences should web archives anticipate and what does this mean for selection, ingest and preservation? What will the web be like as an historical source, and what use will be made of archived web sites by future generations? How will they validate them? How will they cite them? What are our missing links? How can these be filled?

The challenges of web archiving have long been recognised and there are a number of tools and services that already offer – or purport to offer – long-term access to web content. But gaps remain in policy, expertise and implementation and the tools for web-harvesting need a clearer link between the technical needs of preservation services and the deferred needs of user communities.

Only by developing and strengthening the links between content creators, tools developers, preservation services and users can we hope to secure an enduring web.

Sponsored by the Digital Preservation Coalition **(DPC)** and the Joint Information Systems Committee **(JISC)** the six partners of the **UK Web Archiving Consortium** (British Library, National Library of Wales, JISC, Wellcome Library, The National Archives and the National Library of Scotland) organised a joint workshop on the 21st July 2009 at the British Library Conference Centre, Euston Road, London.

This event attracted key stakeholders – archive managers, preservation experts, national libraries, web archivists and content providers - for practical and focussed discussion on shared perspectives, requirements, problems and solutions. Formal presentations and case studies will be presented with an opportunity for posters and demonstrations of tools. The day closed with a plenary discussion.

The conference was free of charge to participants.

2. Feedback and key performance measures

- In total there were 119 attendees on the day. 128 registered in advance. 38 evaluation forms were received.
- Attendees largely represented audiences from national libraries and museums, academic institutions, government, business and industry. Academic institutions and charities accounted for the majority of attendee organisational representation.
- Job roles predominantly included: librarians, archivists, web archivists and managers and record managers.
- The organisation of the event was widely applauded and achieved high targets in all aspects of the following: content of sessions, structure and pacing, venue and catering, organisation and logistics, administration and communication.
- 50% of the feedback forms received from delegates indicated that they were either already or planning web archiving activity on the coming 12 months.
- When asked '*what else would delegates like to have seen covered*?' there was a significant number of comments requesting information on the operational and practical process of web archiving, including actual tools, archiving and preservation, costs and the current state of legal deposit for the web. Further requests included updates on web archiving initiatives in the UK, especially in the politics and social sciences and humanities.
- The final roundtable Q&A session was particularly useful and informative to delegates and at least one delegate would have appreciated an introductory talk on the work of the DPC. Many delegates felt that the topic warranted a two day event and a significant number of requests were noted for a 'follow up' event.

3. Lessons Learned

Speakers were asked to identify the key lessons or messages that they had taken from the conference.

Thomas Risse of L3S Research Centre said:

```
... It was very interesting to see the different views and problems of
Web archiving. Two issue seems in my mind more important now: (1)
selection of web pages (2) integration of archives. Due to the large
amount of web pages it is not possible to have a complete overview
about what is relevant. Therefore we need selection technologies that
decide "intelligent" based in the content if a page is relevant for a
crawl. Furthermore as the crawling can be distributed among several
organisations we need advanced method for integrating crawling,
detecting duplicates or near duplicates. We should also think about
ways to make use of the near duplicates to improve the archive quality.
```

Amanda Spencer of the National Archives said:

```
...I think for us it provoked further consideration of our users - both
current and future users. As immediate concrete steps we are thinking
about how we can do some further curation of our current collection to
pull out areas of interest such as government blogs (Richard Davis's
```

presentation got us thinking in this way). In terms of future users, one of our TNA colleagues from our Digital Preservation dept suggested that we work with the TNA education department to get school children and A level students thinking about the issue of the ephemeral nature of the web and the value of both archiving it and ensuring continued access through links persistence.

Networking with other potential collectors of web archives during the breaks, it's also given us food for thought on how we make web archiving more inclusive, by facilitating take up of web archiving services by smaller organisations.

Eric Meyer of the Oxford Internet Institute said:

For me, I think that Kevin Ashley's question about "what do we want from web archives" is the central question for me. I tried to allude to that a bit in my talk, but I think Kevin did a brilliant job of challenging the web archiving community to envisage what sort of things the users of Web 8.0 will want to know from the archives of Web 1.0, 2.0, 3.0 and beyond.
Extending the debate beyond "what existed in 1999" to "what would have happened if I had searched for this topic in 1999" and "in what context did this thing exist in 1999" and "how did these things on the Web evolve from year x to y" is very important, and a big challenge.

Ed Pinsent of the University of London Computer Centre said:

...I am still reflecting on what was said (I think by someone from University of Edinburgh) about recording and registering the details of IP addresses. This appears to be a dimension of website ownership which has never occurred to me before in the context of web-archiving work. It sounds like it could be very worthwhile for a web archive to be collecting (and updating) that metadata, and it sits well with my traditional archivist's instincts about 'provenance'.

Adrian Brown of the Parliamentary Archives said:

Some of the points which stuck in my mind were:

- How do we bridge the gap between what we want to capture, and what resources, technology and, above all permissions, allow? Can we be less risk-averse?

- Can and should we make a distinction in our approaches between the needs of those who care about capturing specific content (and are most concerned about quality and depth), and those who are interested in trends and the nature of the Web itself (who require breadth and volume, but aren't necessarily concerned about individual content)?

- I guess this one may be answered when we have an announcement about the future of UKWAC, but I think there remains an open question about how we coordinate our collecting activities in the UK, especially beyond the big national bodies.

- Related to this, how we do reduce further the barriers to entry for the smaller organisations - either we need to have affordable services they can use directly or, if national bodies will collect on their behalf, or at their nomination, we need to find ways to ensure that this can be followed through. I think there is a danger at the moment that we get lots of people excited to nominate sites for archiving, and are then unable to fully deliver.

- How do we create the global, virtual web archive?

- Finally, I liked Kevin's point that, rather than trying to second-guess all future uses, and build access systems to meet them, we should concentrate on good, simple APIs which others can build on.

Cathy Smith of the National Archives said:

... more than happy to share a few of my views on the event, which I think was a real success. I was initially concerned that the day was too packed and that such a 'quick fire' programme of twenty minute slots was going to be a challenge. In actual fact, it made speakers focus their thoughts and the result was a series of presentations which covered an amazingly wide range of issues and themes in a relatively short space of time!

As for what I found significant or thought provoking, I'd say that I came away feeling confident that web archiving has finally 'arrived'! The workshop was attended by representatives from encouragingly diverse sectors who helped stimulate discussion and there was a genuine feeling of collaboration and desire for collective behaviour. The future of web archiving is still very much one of learning and experimentation but it seemed that all stakeholders are relishing that thought and are certainly not weary of the challenge. For me, I'd like to see the 'memory' institutions looking to content creators and web developers to help further exploration on the 'how' and to researchers and users to help further debate on the 'why'. I agreed totally with Kevin Ashley's suggestion of providing the technical means for accessing, displaying and mashing-up the data held in web archives ... but believe that he's simply describing a new form of information retrieval which can help augment institutional selection and curation processes.

Jeffrey Van der Hoeven of the Koninklijke Bibliotheek said:

Most remarkable outcome in my opinion is the need for cooperation and alignment of activities which become clear specifically by these kind of events. Since the Internet Archive did start archiving the domains, many newer (small and large) web archiving initiatives have started. Although the IIPC and UKWAC do stimulate dissemination of procedures and tools for archiving the web, there is still a lot of uncertainty about what to archive and who's responsibility it is. The need for a forum and coalitions is more active than ever I think. And a good role for DPC and related platforms as well.

Furthermore, as Kevin Ashley stated very clearly, the web is more than just the web. It's a major source of all kinds of data. If we're able

to preserve not only the content, but the context and computer environment (browsers, plugins, etc.) as well, web archives will be of much more value in the future then we think of today.

Richard Davis of the University of London Computer Centre said:

... even preparing and reflecting on the presentation has useful lessons.

Personally I was particularly struck by the synergy between some of the ArchivePress aims and those of the DACHS project - which I wasn't aware of - relating to citation and persistence of web materials referenced in other work, especially academic research. This also turns to some extent on the transactional model of web archiving that Adrian referred to, and which I find particularly interesting. It's always reassuring to have at least some of one's ideas validated, in what can seem a complex field, with many competing perspectives and priorities.

William Kilbride of the DPC said:

...For my part I have a clearer understanding of how context ought to drive curatorial concerns of content. I mean the observations about the DNS services for example: currently well out of scope but actually very important. By analogy, most of the conversation about early forms of literacy are not about what is written, but who said it and in what context they were able to construct monitor and maintain literacy practices. Who cares what the inscription says - how did it come to be and how was it understood are as important, often times more important.

Extending this thought I'm impelled into a surprising and more urgent concern with archiving and preservation too. Web archiving is JUST PLAIN ARCHIVING: it needs to take place within the wider context of long term digital asset management.

4. Products and value-added deliverables

Each of the presentations and a selection of photographs from the day are available online at: http://www.dpconline.org/graphics/events/090721MissingLinks.html while a larger set of print quality photographs is retained in the DPC office and has been distributed to organisers.

The conference has been reported and content discussed in the following fora outside of the conference:

Clark, J 'How to Archive the Web' in *A wheelbarrow full of surprises (23/07/09)* online at: http://jonathanclarks.blogspot.com/2009/07/how-to-archive-web.html (last access 05/08/09)

Eveleigh, A 2009 Missing Links: The Enduring Web, Ariadne 60, online at: http://www.ariadne.ac.uk/issue60/missing-links-rpt/ (last access 07/08/09)

Guy, M Releasing the Herds of Cows: The Missing Links Workshop, in *JISC Powr (22/07/09)* online at: http://jiscpowr.jiscinvolve.org/2009/07/22/releasing-the-herds-of-cows-%e2%80%93-the-missing-links-workshop/ (last access 05/08/09)

Document Distribution Note
Release to Participants: Immediate
Release to Members: Immediate
Release to Public: 01/09/09

Winters, J and Webster, W Report on 'Missing Links: the enduring web' conference' in *IHR Digital (05/08/09)* online at: http://ihr-history.blogspot.com/2009/08/report-on-missing-links-enduring-web.html (last access 05/08/09)

Author/Editor	**Date**	**Action**
WK	05/08/09	Document initiated
CJ	07/08/09	Feedback highlights added
WK	07/08/09	Release to participants
WK	07/08/09	Release to DPC members
WK	01/09/09	General Release