Release to members Immediate
General Release 10/01/2012

DigitalPreservationCoalition

Report from 'The Future of the Past of the Web', London 7th October 2011

Matt Brack, King's College London

1. Introduction

This document reports the one-day international workshop 'The Future of the Past of the Web' held at the British Library Conference Centre, London on 7th October, 2011. Co-organised by the DPC, the British Library and JISC, this workshop was the third in a series of discussions around the nature and potential of web archiving. Following the key note address, two thematic sessions looked at 'Using Web Archives' and 'Emerging Trends': it is only recently that use cases have started to emerge, while it is acknowledged that web archiving activities are on the increase, along with a corresponding rise in public awareness. In addition, this report serves to assist DPC in measuring the success of its events programme, help shape future work in the field and provide a commentary for those DPC members unable to attend. It includes a narrative report drafted by Matthew Brack, who has also compiled evaluation from participants.

2. The Future of the Past of the Web

The Web expands at an astonishing rate. Statistics suggest that more than 70 new domains are registered and more than 500,000 documents are added to the web every minute. This rapid expansion continues to challenge those charged with preserving an effective memory of the web.

Memory institutions – in particular national libraries and archives – have been central to web archiving. Since the mid 1990s, they have captured a dynamic and highly distributed snapshot of the web as it evolved. These growing web archives provide an untapped resource for creativity, innovation and enterprise. The web archiving community has grown as more institutions establish their own web archiving programmes. Universities and researchers are also taking part in this effort and commercial archiving services have started to appear.

Use and impact of web archives are under-explored topics in discussions about web archiving. Alternative modes of access and new types of exploitation mean that the time is ripe for another examination of how the web archive collections are being used and what opportunities they open up. Web archives are no longer just individual web pages for reference but also aggregated datasets with inherent properties which can be exploited for many new possibilities.

3. Narrative Report

- **Key note (Herbert van der Sompel, The Memento Project)**

The Memento Project: The Web is a communication device, with resources that change over time. The Web was built with the notion of the 'perpetual now': at any moment in time, it is only possible to obtain the current representation of a resource, and it is not possible to get a past representation of the resource through a URI. Some content management systems can give past views of resources, such as web archives, or memory caches from search engines, but these are not integrated into the Web itself. For example, when you do find an old resource, you cannot navigate in the past – links take you to current versions of a resource, rather than providing the contemporary context. The Memento Project seeks to tackle this problem by bridging the gap between an original resource (for which we want to find a prior version) and a prior resource (or 'memento') by introducing a 'TimeGate' that provides an intermediary resource linking them at the level of HTTP protocol

between present and past. It actually works, using a plug-in for your browser that allows you to select a date from time.

Web impact on the scholarly record: The characteristic of publications from the paper era are that they are frozen in time, but this has changed with the advent of the Web. What we are citing now is not only a paper frozen in time, but also web resources – they both exist within a broader web context. Web content changes, and the material that sits around a publication certainly changes – all of this has the disease of the 'eternal now'. You only see a publication in its current context and not in the context in which it was originally published. Is it possible to recreate context from the time of publication? In fact we can't actually recreate the context, because we are not archiving the context, only the papers themselves. DOI (digital object identifier) can provide redirection for link persistence with documents, but redirections change and new resources come up. If you want a 'memento' of a paper from the past, the browser only sees the current redirection and not the previous resource we wanted. You therefore need a TimeGate for the DOI itself – from that you can follow the TimeGate link to the right resource and the appropriate memento. In conclusion, we should be looking beyond the papers themselves and towards the context which surrounds them, and this is crucial for examining the scholarly record.

Thematic session: Using Web Archives (chair Helen Hockx-Yu, British Library)

- **Web Archiving: the state of the Art and the Future (Eric Meyer, Oxford Internet Institute)**

While there is hope that people will be using web archives on a grand scale, there is also reason to be dissatisfied with the persistent gap between researchers and the creation of web archives. There is a fear that web archives could become the 'dusty archives' of the future. This is ironic, since so many of the dusty archives of the world are now being moved onto the Web. How can we avoid web archives becoming dusty archives? What steps can be taken for their use by researchers? Social science researchers don't seem to engage with web archives, and when asked about web archives they don't understand them. This is a challenge, because it's always more easy to do more of what's already been done. We gain disciplinary bias early in our careers, so that tends to discourage innovation – like trying to understand using web arches, which is not part of that acquired scholarly behaviour. Yet there are massive events going on in the world – if a scholar is interested in collecting evidence on the web for what's happening, most people don't know how. They don't go to libraries because they don't know what libraries do, or what's available. We need someone to help people archive this material on an individual basis and develop ways to trigger collecting. Server logs aren't saved – what was happening with a resource 5 years ago is lost. Could logs be donated to web archives for analysis? The most ambitious goal is archiving web traffic itself, though obviously there are legal implications.

- **Enhancing Access to Web Archives: Web Continuity at the National Archives (Amanda Spencer and Tom Storrar, The National Archives)**

The National Archive started archiving UK Government websites in 2003. At first this provided only a 'snapshot' of Government Web publications, but it now includes all government departmental groups, public enquiries, royal commissions, and even NHS websites. These websites are prone to link rot like any other, and parliamentary librarians soon noted that URLs cited in UK Government publications could no longer be found. The National Archives then developed the Web Continuity Initiative for link persistence within important government information and began employing software to redirect users to the UK Government Web Archive (Archived websites can also be discovered through a Google search). Content loss is more problematic during 'machinery of government' changes, particularly after general elections. Current projects include: a secure web archiving project with Hanzo Archives, with a view to ultimately providing public access; and social media capture, including YouTube videos, Twitter feeds and Flickr content. Having identified the

limitations of their current conventional search solution, a semantic search is being developed. This will structure unstructured content by assigning meanings, so the system can understand what these concepts are. Information curated around that can then define relationships between these concepts. The key to this project is to build something useful. It will launch next year with a user interface for non-technical users, and an API for developers.

- **Case Study: Researchers and the UK Web Archive Project (Peter Webster, Institute of Historical Research)**

A team of scholars were assembled by the British Library to guest-curate a collection in the UK Web Archive. Collection subjects included independent artist-led organisations, digital storytelling, the politics of religion in the UK since 7/7, and gender, power and the Olympics. The curatorial aim is to try to create a resource for someone 50 years from now. Websites of interest are those which attempt to represent previous ages, and touch upon history and the media (for example, a website covering thoughts on the anniversary of the King James Bible). Website of interest also include those which illustrate major events, new developments in national life, or older movements that are ending. For example, the question of whether there should be bishops in the House of Lords has generated multiple websites expressing differing views. The implications of legal deposit are significant: if new legislation is passed next year this will remove permissions issues. Currently, only 25-30% of requests for archiving are successful. This means that desired Web content could disappear before legislation is passed. Legal deposit would mean more comprehensiveness, but the focus then shifts from selection to curation. There is an opportunity to use the crowd – the social media tools are there for allowing users to browse and tag materials interesting to them. The missing link is that there is little in the way of any agreement to include content from the press in the archive. One hopes that someone archived the News of the World's website before it disappeared.

Thematic session: Emerging Trends (chair Neil Grindley, JISC)

- **Analytical Access to the UK Web Archive (Maureen Pennock and Lewis Crawford, British Library)**

Data mining is on its way. Real user value means data level exploitation of web archive content. The UK Web Archive was first explored during 2001-2002, finally becoming fully embedded as an operational unit at the British Library this year (2011). An interface from 2005 shows former page level access. The new interface now has more access options, including a 3D visualisation wall (particularly useful for assessing overall webpage design), word cloud access to special collections and an N-gram search function for instances of terms over time. By extracting the metadata behind images, there is also an enhanced image search, displaying in the manner of Google Images. Collections are curated before they are collected, so they are broken down into various subject headings. It is a relatively small archive, containing only 1% of UK domains, but already it would take 3 hours to read through all of these headings. A data mining experiment has mapped post codes by identifying 42,000,000 relationships between web pages and post codes. This can then show density of post code records across the UK. If, for example, they were filtered by a category like 'business' and some associations disappeared between 2007-2009, could it be possible to map the recession?

- **Emerging Trends and new developments at the Internet Archive (Kris Carpenter, Internet Archive)**

The Internet Archive is interested in the direction we are going with the infrastructure, data, questions to ask of resources and collaboration with communities of interest who want to leverage the Internet Archive collections. The goal is to enter a body of material, manipulate it and delve into it as something you can access from your living room or place of work, linking a diverse set of resources in more meaningful ways. We have been focusing on how content is published, but it's

also accessed in many different modes. Two trends have developed: first, the personal domain of digital content and archiving – individuals produce 70-80% of content that's out there; second, everyone at some point has a need to pull all the resources they use together. How do we address access using diverse modes of interaction and rich data sets? How do we re-create the experience the user had at a certain point in time? The Web is a mess and even identifying the subject of a resource in automated fashion is hard. The Internet Archive are trying to develop scalable architecture to mine data on a large scale to extract metadata to get clues about the content within a resource. We've talked about the 'hidden' Web for a decade – the challenge is that it's expanding, now more than before. There's an interest in taking a wide range of data types and making them available, but how do you make them a part of the broader ecosystem in a meaningful way? In the case of social networking, if there is nothing that can archive that context and usage, how do we represent it to someone examining that social interaction 20 years from now? Collecting this content is easy – the bigger challenge is re-rendering it from an archive to show the original resource it was referenced by. Traditional crawlers are effective in content collection, but you need a hybrid architecture that shows you are getting all the elements that are needed to re-render a resource – most resources are made from 35 different files. Working at domain scale over 10 years of history we are looking at what we can discover from the evolution of that content and have identified 200-300 million unique entities that we might want to represent as aggregations for study.

- **The Arcomem Project: intelligent digital curation and preservation for community memories (Wim Peters, The University of Sheffield)**

There are a number of challenges in archiving the social web: networks are ephemeral, there are many communities, there is uncertainty of future demand, and the technical challenges involved in capturing this information are significant. You can't rely on a collect-all approach, it needs to be filtered and measured against criteria of demand: community memories that reflect communities' interests. What is a reasonable way to create collective memories? What is valuable content? Arcomem is trying solve this by relying on crowds for intelligent content appraisal, selection, and preservation, with collaborative content acquisition support for archivists. Collective memory, as well as personal memory, revolves around topics, events and entities. The overall aim is to create an incrementally enriched web archive that allows access to various types of web content in a semantically meaningful way. The main challenge is to extract, detect and correlate events and related information from a large number of heterogeneous Web resources. Arcomem collects with intelligent crawling, focusing on events and entities for detection and extraction. Archivists can select content by, for example, describing their interests by using reference content, or by a high-level description of relevant entities. Finally, Arcomem seeks knowledge consolidation by creating common reference points for data extracted from different components and resources, embedding archival info objects into a wider network of knowledge. Arcomem will be interoperable with external resources with linked open data, event and entity models.

- **BlogForever Project (Richard Davis and Ed Pinsent, ULCC)**

BlogForever is an EU-funded blog archive. Things continue to change and evolve on the web and some might consider blogging 'old hat', but it would seem that this particular communications paradigm is here with us to stay. You can preserve a blog page, or even print it out, like the paperback version of *Geoffrey Chaucer Hath a Blog*, but it certainly lacks something – somehow it's not the same as it is online. So we may have the content but not the same context. The objective of BlogForever is to develop robust digital preservation management and dissemination facilities, while capturing the rich essence of blogs (unlike a printout). The outcomes are the definition of a generic data model for blog content metadata and semantics, as well as the definition of digital preservation strategies for blogs. Unstructured blog information is collected and then given a shape that can be interrogated in all sorts of ways. The project has only just begun, with the structure and semantics completed thus far, while work continues on the policies. A survey was conducted of bloggers'

attitudes to preservation: 90% never used an external service to preserve their blog, but relied on the blog provider for preservation; 30% used re-mixed data, so this could raise permissions issues. The result will be a weblog digital archiving solution.

- **Web Archiving: a commercial perspective (Mark Williamson, Hanzo Archives)**

There are several companies offering commercial archiving services, indicating a commercial interest in web archiving. The web is the first media in history that has no way of being 'kept', yet it is where the majority of human creation is happening now. Commercial organisation are now worrying about research problems. Some companies are forced to archive web material because they are regulated, and many of their webpages are extremely complicated. We've heard lots about social media – who would have thought that banks and large corporations would be worried about their Twitter streams? Social media is structured data, it's a conversation, it's posts – they also want to discover things. Companies have large amounts of digital data and are interested in discoverable content, such as trying to pull out social identities (who wrote what and so on). The new thing emerging in the last few months is that people are coming to commercial archives wanting to collect not just for the content's own sake, but because they are interested in data and scale. In the world of business archives, companies are worried about the name 'archive' – they would prefer to call it 'high resolution data', or some such thing – for a lot of people it's an old word, a dead word. In many cases the individual pages are pretty dull, but it's about the big data. There is a huge tide of this data, and there are not enough archives to collect it. Hanzo will be opening up tools via an API to bring web archiving to everybody in the near future. This API will let you crawl on demand.

Panel session and discussion: what is to be done, why and by whom? (Chair William Kilbride)

What is to be done?

- Are web archives heading in the right direction?

  In many ways the expectations held during the last discussion two years ago have been exceeded – what we want to do we can do, in some cases we have gone beyond. Thinking of the web as data can help to address the problem that we can't keep everything. Perhaps in some cases we can afford to just keep the wrappers around content, showing us what links to what, rather than exactly what was said. Is there a role for these stripped down cases? Access isn't about a new user interface, it's about prompting different means and methods for access.

  Martha Anderson from the Library of Congress commented that *no one* can serve up the whole of Twitter (not even Twitter themselves), and so of course the whole of the internet will not survive. Web pages as a medium are already dying and the idea of a web page is similar to our idea of a bound book – we find it very hard to imagine something we are so familiar with is changing in front of us. This revelation came for the Library of Congress during a project to document the Japanese tsunami. They found that it was the social media that told the story, not webpages constructed to memorialise the event. What people are interested in is getting finer and finer. In web archiving, just as the river shapes the land and the land shapes the river, our practice is shaped by the media we are dealing with. That changes and we change as well.

By whom?

- Why hasn't anyone mentioned Google?

Google has been able to preserve their original search engine code. During their 10th anniversary they wanted to deploy that code, which pointed into the Internet Archive. Computer scientists in the US are also interested in working on the now defunct Altavista search code. There is certainly confusion in the user experience within web archives, especially with different versions of the same documents. The technology that Google has already displayed to help put search results under one heading would be better, rather than sending users straight into the archive and getting lost. At that level one could use search engines more to access archives – you need a uniform way of de-duping those things. Essentially, we're not asking for a huge leap to take place, the technology is already there. Though search engines play a role in search, they don't play a role in web archiving. You can't use their cache programmatically. They could play a larger role, but don't at the moment.

- Does the decentralised Web mean more individual archiving?

We've been saying how websites don't really exist anymore. While it's decentralised on an individual level, has it centralised on a community level? People used to have personal websites, now people have blogs that serve as an aggregation facility for many of them. When we started archiving we were collecting pages, because we thought people would be looking at pages. We also found out when researchers came, they weren't interested in pages and pictures, they brought scripts on pen drives that they ran across the aggregate. We see this accelerated now with apps on portables because the webpages do not exist anymore, the data is pulled from them and delivered instead. With so many more people creating content now, does this mean that they will be more concerned about archiving? The BlogForever survey said that 90% of bloggers thought their blog provider would preserve their blog. What do we say to them? We have said: "You realise your blog won't be preserved at the moment," to scare them, and now we want to focus on the positive aspects.

- How do we bring in these users?

The digital preservation programme at the Library of Congress found that a couple of years ago it was fine to convince congressmen that this was important but they needed grassroots advocacy – so now they have personal digital archiving. They have talked to thousands of people, often focusing on personal photographs, saying, "You need to think about what you really care about and take some measures." Perhaps some day we'll see posters on buses that say, "Preserve! Are you saving your digital stuff?" But there is a cultural barrier: many people don't think this is important enough to bother anybody with. Personal digital archiving is probably the way to go – one person at a time getting the message out.

- Should we approach providers?

Is there a gap here regarding Google, or YouTube, or Flickr? All it takes is for Flickr to switch off and everything disappears. Should we be doing more in that space, too? One approach is that you go to providers, the other is that you go to individuals and say, "If that's where you're storing it, in Flickr, it's not safe." It's been found that during a presentation of these issues to 8-9-year-olds, they understood the risk to digital content, and the awareness was there. At one point there were some companies who had been targeting the individual blogger – offering a permanent archive of their blog for a one-off sum of money. This was actually patented. Are those companies still here now? And did they make any money? There are some changes in private sector thinking about these issues, because the consumer

has started wanting to archive. Wordpress have been interested in developing tools that allow a user to ask for their blog to be archived to a specific repository. Flickr and Yahoo have started to talk about this, too. The time is right to take leadership in this community – not forcing archiving but presenting the option.

Where do we go from here?

- Having worked on this for ten years, what should we now focus on?

The organisation Global Voices, who collect citizen blogs from countries involved in the Arab Spring and events all over the world, visited the Library of Congress . They were concerned about the content. It's a curated collection – they are not librarians or archivists, but they have a concern. Those are the opportunities: hundreds of blogs in one collection that don't need to be looked at – people should take advantage of the community out there, and have a broader approach. Rather than using a single interface for the whole Web, we could get communities to devise bespoke methods for access. The bridge between institutions and communities is not as strong as it could be.

- Are librarians happy with tools for automated selection?

Web services that allow mining exercises are great. This is not to advocate not keeping traditional navigation, but data mining will be really big. At the first DPC discussion on this in 2002, there was already an algorithm better than human selection. Now we have far better technology, yet there's a feeling that somehow we can't quite trust this in libraries and archives. It's not that there is no role for humans, but we have these technologies, and in ten years not that much has changed – it is still not accepted by librarians.

- Do we need standards?

We've heard about the grand challenges and the divergence in how we need to work, but what do we do about standards? There are lots of ways of doing this, should we think about convergence? We've got WARC format, do we need more? Both ISO and BSI have been working on things, putting together a report that could be used to assess web archives, though probably more from a service provision point of view. It's work in progress.

- What are the next steps?

There is a lot of room for optimism and we've seen a lot of examples today. The next step is to get such examples into communities that can imagine uses for them. Time and again, the best way to communicate this is for peers to present to each other within the same domain. It's difficult to just be shown these tools, but if you present it to historians as a way of answering a historical question, for example, then there may be more success. Those are the next steps – it needs to percolate out to those who will dream up lots of clever things to do with these archives given the right tools and incentives.