LIBER QUARTERLY 18 (3/4), December 2008 – ISSN: 1435-5205. P357–365
http://liber.library.uu.nl/
Igitur, Utrecht Publishing & Archiving Services

Web Archiving in the UK: Cooperation, Legislation and Regulation

John Tuck

[at the time of writing] Head of British Collections, The British Library,
96 Euston Road, London, NW1 2DB, United Kingdom

Abstract

The author presents an overview of web archiving in an international context, focussing on web archiving initiatives in the United Kingdom from 2001 onwards.

Key Words: Web archiving; UKWAC; United Kingdom; legal deposit

The Beginnings

‘The Web today plays a crucial role in our information society: it provides information and services for seemingly all domains; it reflects all types of events, opinions, and developments within society, science, politics, environment, business, etc. Due to the central role the World Wide Web plays in today’s life, its continuous growth, and its change rate, adequate web archiving has become a cultural necessity in preserving knowledge.’ These words are taken from a recent description of the Living Web Archives project in Digital Preservation Europe’s *European quarterly preservation digest* (July 2008) and form a solid rationale for the substantial web archiving activities which are taking place worldwide.

Web archiving really began in 1996 when Brewster Kahle had the vision of collecting the universal web. Brewster Kahle is described by wikipedia as a ‘U.S. internet entrepreneur, activist and digital librarian.’[1] He was the founder of the Internet Archive (IA) whose goal was and is to build an internet library.

Pioneers in the field and collectors, in 1996, of what some have described as the incunabula of the web were the Internet Archive, setting about global collecting, the National Library of Australia which set about selective harvesting and the National Library of Sweden that embarked on a crawl of the Swedish domain.

At the time there was considerable need to justify why this collecting was necessary. To many it seemed like a technology — rather than a collection-driven activity. Now, however, as articulated in the very first words of this paper, the why question no longer needs to be asked. More specifically, and for those who do require any justification, the reasons include the transient nature of web pages; the web as an integral part of our lives; the unique material only available on the web; the loss of content as we speak and write; and the role of national libraries to collect, preserve and make available material in all formats.

Early Collaborations

It soon became apparent that web archiving was an area of activity where collaboration and cooperation were required to address the complex technical, legal and operational issues. Cooperation began around 1997 through discussions at IFLA; through the collaborative work and thinking of the Nordic countries; through the working together of the Library of Congress and the Internet Archive.

In 2003 the Conference of European National Librarians (CENL) held an important meeting in Paris. As stated in the CENL annual report 2002/2003:

'A special meeting of CoBRA members and representatives of the Library of Congress and representatives of the Internet Archive (IA) took place in Paris in January 2003 to discuss a proposal on web archiving from the IA. Follow-up meetings also involving the National Library of Canada and representatives of Nordic national libraries have taken place. Representatives of this group will form an informal consortium to work with the IA initially on the specification and development of open source web crawler software.' (van Trier, 2003).

The cooperative and collaborative nature of web archiving really became formalised in 2003 through the establishment of the International Internet Preservation Consortium (IIPC), whose remit is to:

- enable the collection, preservation and long-term access of a rich body of Internet content around the world;
- foster the development and use of common tools, techniques and standards for the creation of international archives;
- be a strong international advocate for initiatives and legislation that encourage the collection, preservation and access to Internet content;
- encourage and support libraries, archives, museums and cultural heritage institutions everywhere to address Internet content collecting and preservation.

Web Archiving in the UK: the Beginnings

The web archiving journey in the UK had, however, already begun. In 2001 The British Library, through Dr Stephen Bury, carried out an in-house pilot, Domain.UK, seeking permissions for and harvesting approximately 100 websites. This Domain.UK work was a spur to further developments as were two key reports, the first produced for the Joint Information Systems Committee (JISC) and the Wellcome Trust in 2003, called *Collecting and preserving the world wide web* (Day, 2003) and the second, focussing on legal issues, by Andrew Charlesworth, reader in law at the University of Bristol, called *Legal issues relating to the archiving of internet resources in the UK, EU, US & Australia* (Charlesworth, 2003). It was really due to the Domain.UK pilot and these two reports that the British Library's web archiving programme and the UK Web Archiving Consortium (UKWAC) came into being in 2003.

Developing Collaboration

It is interesting to reflect on the reactions of colleagues as we have progressed collectively along this interesting and challenging road. 'Scary', said Richard Boulderstone (Director of Information Strategy at the British Library) at the start. 'A bumpy ride' and 'How do you eat an elephant', said Thorsteinn

Hallgrimsson (National and University Library of Iceland and Chair of the IIPC Steering Committee). 'If something is worth doing, it is worth doing badly', said one of the Australian pioneers. In other words, when something is important, be prepared to take some risks, to learn by mistakes and by trial and error.

In many ways this is what has been done through working cooperatively in the UK through UKWAC and internationally through the IIPC which at present has 37 members (including Japan and China). The IIPC is now mature, focussing on harvesting, preservation guidelines, indexing and access, and developing standards for web files. Worldwide, billions of files have been collected and petabytes of storage filled. The agenda is still very much on developing tools to facilitate the web archiving process and to address the technical challenges posed by new software, by flash, video, film content etc., but it is also much more on what researchers want and how they will use this vast research resource which national libraries and other institutions are collecting. 'Imagine what users will do with the web archives in 2015' is, for instance, the title of a session at the September 2008 IIPC webstorming meeting in Arhus.

Web Archiving in the UK: UKWAC

UKWAC was set up in 2003 and comprised six institutions: the British Library, as lead partner, the National Library of Scotland, the National Library of Wales, the National Archives, the JISC and the Wellcome Trust. The project objectives, as set out in the agreement between the partners, were to:

- carry out a two-year pilot web archiving project with UK partners, focussing on (but not necessarily limited to) the UK web domain;
- establish a collaborative enterprise with shared objectives and costs;
- develop a web site to allow archived resources to be accessed via the internet;
- undertake an evaluation exercise to help determine the long-term feasibility of web archiving.

The objectives were fulfilled and, further to the evaluation, a two-year extension was agreed. UKWAC operated, where appropriate, on a rights-cleared basis. A common permissions form was devised and only when in receipt of explicit permission from a website owner, were sites harvested on a selective basis.

In the case of the British Library the success rate of the permissions process has been somewhere between 25% and 30%, although there have been very few refusals. Statistics to the end of March 2008 show that the British Library had selected 6,609 sites, had sent 7,476 permissions requests (including some reminders). 1,872 permissions had been granted and there had been 51 refusals.

To run the system and the workflows, UKWAC used the PANDAS software, developed by the National Library of Australia and kindly licenced to the Consortium free of charge. A third-party supplier, Magus, hosted and supported the service.

Once harvested, the sites were made available free of charge through www.webarchive.org.uk where, at present, 2,769 sites are available. They reflect the collecting policies of the six libraries and include a number of themed and topical event collections, for instance, general elections, the London bombings of 7 July 2005, and several others.

The numbers are small as is the storage space occupied, just less than two terabytes currently. This can be explained by the limitations of a permissions-based approach in a legal environment where whole-domain harvesting is not yet possible. It should also be noted that a learning curve was required at the outset of this collaborative venture and the PANDAS software did have its own limitations, especially when applied to the collaborative environment.

At the end of the agreed UKWAC extension period, September 2007, a decision was reached to move to a new infrastructure and new tools, taking advantage of the developments made in web archiving through the IIPC. A new British Library-led service is in the process of implementation and will be fully up and running with four partners in August 2008. The new system utilises the Heritrix crawler and the Web Curator Tool developed by IIPC, and for which the British Library is the current lead partner for coordinating developments and new releases under the auspices of IIPC. The archive, which will continue to be branded as www.webarchive.org.uk, is being hosted by the University of London Computing Centre pending migration to the British Library's digital object management system. The new infrastructure and tools will be capable of whole domain crawling, the implementation of which is inextricably linked to the e-legal deposit implementation process in the UK. The National Library of Scotland continues its web archiving programme through the Danish NetarchiveSuite system.

Web Archiving in the UK: the Universe

For purposes of supporting a case for whole domain crawling under legal deposit for free UK online publications, calculations have had to be made about the size of the UK web universe and the size of websites. A mass of data has been collected and analysed and has led to the following estimates. There are currently more than 6 million registered UK websites and the trend of registration seems to suggest an annual growth of 16% in the number of sites. The assumption here is that the UK web space is defined as all .uk domains registered by Nominet, the official UK registry, plus approximately 50,000 other domains which can be readily identified as published in the UK.

Calculations were also made to assess system capacity and capabilities. These led to an estimated average website size of around 25.4 megabytes, this average size growing by about 5% each year. This estimate includes images and photographs but excludes most audio and video content as that is out of scope for UK legal deposit. The figures also only apply to the UK free web to fit in with the recommendations for legal deposit regulation as described below.

These estimates demonstrate that a selective permissions-based approach is not sustainable if the goal is to fulfil a legal deposit obligation of acquiring comprehensive UK web content. The calculations across the six UK legal deposit libraries suggest that continuing with a permissions-based approach would result in acquisition of less than 1% of the UK domain after ten years and the library costs of doing this, per terabyte, would be about £6,500. A regulation-based approach, based on legal deposit legislation and not requiring permissions would be much more cost effective, enabling whole domain harvesting and yielding an estimated 81% of the domain after ten years at a cost of £215 per terabyte.

John Tuck

Web Archiving in the UK: Legal Deposit

The Legal Deposit Libraries Act 2003 reached the statute book in 2003 and formed enabling legislation to extend legal deposit provision to non-print. In order to enforce the Act and for anything to happen under the Act, secondary legislation is required. To progress this, the government set up the Legal Deposit Advisory Panel, known as LDAP.

LDAP came into being as a non-departmental public body in 2005 and comprises fifteen members: five librarians, five publishers and five independent members. Its remit is to advise the Secretary of State on the timing and content of regulations relating to legal deposit and to oversee the implementation of the Legal Deposit Libraries Act 2003.

The focus of much LDAP work during its first years has been on three formats: offline; scholarly e-journals; and free UK online publications. By the latter is meant the visible free web where there is no barrier to access, for example, no e-commerce or subscription barriers.

The process for each of these formats has been an options appraisal. For instance, in the case of free UK online publications, the costs, benefits and disadvantages have been assessed for three options: a) permissions-based harvesting and archiving (the UKWAC model); b) regulation-based harvesting and archiving; and c) archiving left to the market.

Significant work has been carried out to cost and document these options which were presented to LDAP at its May 2008 meeting, together with a recommendation for the full regulation option. The full regulation would allow the legal deposit libraries in the UK to harvest, preserve and make accessible (within restrictions) this category of material without the need for permissions while at the same time affording publishers and libraries protection on copyright and defamation through the Act. LDAP accepted this recommendation from its subcommittee but with the proviso that a number of issues raised by the newspaper industry needed to be clarified and addressed. This dialogue is taking place at the time of writing.

The next step, subject to the above clarification, is for the recommendation to go to the Department for Culture, Media and Sport whose officers will provide the necessary wording for the next stages in the legislative process which are an economic impact assessment and a public consultation. Findings from the consultation will need to be evaluated before the regulation is put to both Houses of Parliament for their affirmation. The earliest time for regulation to be implemented would, therefore, seem to be late 2009 or early 2010.

It should be noted that the Legal Deposit Libraries Act 2003 places restrictions on access, which, in general terms, can be described as access on UK legal deposit library premises only. In other words, unless further permissions are sought, free UK online content, harvested under legal deposit, would only be accessible within the reading rooms of the six UK legal deposit libraries.

In parallel with this legislative process, preparatory work is taking place in the UK legal deposit libraries to ensure appropriate means of receiving, preserving and making accessible this material. Three — the British Library, the National Library of Wales and the National Library of Scotland — are developing a shared technical infrastructure both for preservation and access to e-legal deposit content. The other three legal deposit libraries — Cambridge University Library, the Bodleian Library, University of Oxford and Trinity College, Dublin — do not have current plans to develop their own web archiving systems and infrastructures but intend to access the shared infrastructure of the three national libraries.

Conclusion

Within the overall context of the British Library's more than 150 million collection items and its overall budget of approximately £130 million per year, the web archiving collection and operation still remain small. However, as the journey which is web archiving develops and grows, it soon becomes clear that the web in terms of files and items will soon dwarf printed collections in the traditional numerical sense. The origins of the British Library go back some 250 years but those of web archiving a mere twelve. In that short time significant progress has been made on the web archiving journey. In general, the web archiving community in the UK and internationally knows where it wants to go. We are engaging with the researchers; we have a range of people across many libraries and institutions who have deep knowledge and expertise in this fascinating field. We are working together nationally, internationally and enthusiastically to provide both a system and access to a rich research resource which will run to trillions of items.

References

Charlesworth, Andrew (2003): *Legal issues relating to the archiving of Internet resources in the UK, EU, USA & Australia*. JISC/Wellcome Trust, 25 February 2003): www.jisc.ac.uk/uploaded_documents/archiving_legal.pdf

Day, Michael (2003): *Collecting and preserving the World Wide Web. A feasibiblity study undertaken for the JISC and Wellcome Trust.* JISC/Wellcome Trust, 25 February 2003. http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf

Van Trier, Gerard (2003): *Annual Report Foundation CENL, July 2002–June 2003.*

Websites Referred to in the Text

CENL, Conference of European National Librarians, http://www.cenl.org/

European quarterly preservation digest, http://www.digitalpreservationeurope.eu/publications/newsletters/Quarterly-Bulletin-Final.pdf

JISC, Joint Information Systems Committee, http://www.jisc.ac.uk/

IA, Internet Archive, http://www.archive.org/index.php

IIPC, International Internet Preservation Consortium, http://www.netpreserve.org/

LiWA, Living Web Archives project, http://www.liwa-project.eu/

UKWAC, UK Web Archiving Consortium, http://www.webarchive.org.uk

Wellcome Trust, http://www.wellcome.ac.uk/

www.webarchive.org.uk

Notes

[1] See http://en.wikipedia.org/wiki/Brewster Kahle [accessed 17 July 2008].