
Selective web archiving in the UK : a perspective of the National Library of Scotland within UK Web Archiving Consortium (UKWAC)

Paul Cunnea
*E-Collections Development Co-ordinator,
National Library of Scotland*
E-mail: p.cunnea@nls.uk

The following article describes the National Library of Scotland's (NLS) experience of web archiving within the UK Web Archiving Consortium (UKWAC), and provides a summary of the work of the consortium, which is in the early stages of building a UK national web archive for present and future generations.

BACK IN THE BEGINNING ...

... for the NLS web archiving began at the end of the last century, when the library carried out a small web archiving pilot to cover the 1999 election of the first Scottish Parliament in almost 300 years. This was a relatively modest affair, focusing on the 14 websites of the parties who stood in the election, but with a very similar permission-based approach to that of UKWAC outlined below.

Both the British Library (BL) and the National Library of Wales (NLW) were carrying out similar pilots around this time, with the British Library's 2001 Domain.UK project covering, amongst other topics, the last UK general election – as you will find, parliamentary elections are very popular in web archiving circles.¹

Globally, systematic web archiving has been carried out as far back as 1996, with organisations

such as the Internet Archive and National Library of Australia (NLA) creating publicly accessible web archives, and the more recent Nordic web archive building on the earlier work of the Royal Library of Sweden and the other Scandinavian national libraries.^{2,3,4,5} More recently national libraries around the world have taken up the challenge, with the Library of Congress Minerva project in the US being a notable example.⁶

BUT WHY ARCHIVE THE WEB?

It has been claimed that the average lifespan of a website is 44 days⁷. Some might say that this is already far too long for much of what is on the web, and that 'important' sites will have far longer life spans. However, it is clear that significant content, whether it is social, cultural, political or topical, is being lost from the web on a daily basis. The lack of systematic collecting of web content can be likened to a potential dark age for digital information, with huge gaps in our recorded knowledge of what is being said, created and published in this medium. Even scientific knowledge is being lost, with a recent survey reporting a 'decay rate' of 33% for online citations.⁸ If information is being lost for today's web researchers and browsers, how much more so for future generations, unless efforts are made to maintain and preserve unique, born-digital web content?

THE CREATION OF THE UK WEB ARCHIVING CONSORTIUM (UKWAC)

Owing to this increasing awareness that a 'dark age' of electronic information may result from a lack of action, as well as recommendations from a JISC (Joint Information Systems Committee)/Wellcome Trust feasibility study, the UK Web Archiving Consortium was formed at the end of 2003, and charged with tackling UK web archiving on a collaborative basis.⁹ The consortium consists of the British Library, the National Library of Wales, the National Archives, JISC, and the Wellcome Library, as well as the National Library of Scotland. All partners have an interest in collecting websites for future posterity, and have pooled resources to pilot the UKWAC project.

The project was officially launched in June 2004, following selection of the appropriate software and support environment. As a two-year pilot, the membership remains fixed for the duration of the project, but, assuming successful outcomes, it is expected that the work of the consortium will

continue, and that collaboration will be extended to new partners.

COLLECTION AND SELECTION POLICY

Under this collaborative framework, each member will select websites according to their own collection policy, but will collaborate and coordinate efforts to avoid duplication, increase efficiency, and to provide as wide a coverage as possible for what is effectively a single web archive. The National Library of Scotland is collecting websites of Scottish cultural significance, with culture being interpreted in its widest sense.¹⁰ For the NLS, although selection criteria inform the process, nothing on the web is excluded from consideration: from local community sites and arts festivals, to national organisations and events, nothing is too big or too small.

AIMS OF THE PROJECT

In essence the pilot aims to test the feasibility of selective, permission-based web archiving within a consortial framework, and to build an archive of at least 6000 websites within the two years of the project. The key aims of the project can be summed up in the following:

- to test the feasibility of selective web archiving in the UK
- to test the feasibility of permission-based web archiving in the UK
- to explore the benefits of a partnership approach
- to test the methodology and software used
- to define a sustainable, long-term solution
- to create a searchable and public web archive.

Given the acceptance that no single organisation can hope to archive the web effectively on its own, and that wider collaboration and partnerships will be necessary, finding solutions within the consortia based approach is very important to the project.

THE TECHNOLOGY – PANDAS¹¹

Following extensive evaluation by consortium partners, UKWAC selected the PANDAS system. This is an integrated web archiving management system, and one of the few solutions suitable for the aims of the project. Developed by the National Library of Australia, and available via open source, it has been used by them to build their Pandora web archive (see above) over the last eight years on a similar selective, permis-

sion-based model. The system uses Oracle to hold the metadata, whilst the gathering of sites is handled by the open source web crawler/harvester HTTrack.¹² The consortium has contracted Magus Research to host, maintain, and provide a support environment for the PANDAS system.¹³

THE ARCHIVING MODEL

As detailed above, UKWAC is using a selective approach to web archiving, similar to the NLA model. This aims to archive a selection of websites within UK web space (and in some cases outside the UK domain), in line with partners' collection policies, and does not aim to be comprehensive. This is opposed to whole domain archiving – as carried out by the Internet Archive and others – which aims to archive everything within a particular domain or area of the web, e.g. all sites within the '.uk' domain, or indeed all sites within the web.

It is generally accepted that neither approach is adequate: selective archiving by its very nature will not be comprehensive, whilst global/whole domain archiving is unlikely to achieve the depth or frequency required for archival quality collections. A combination of approaches is likely to be necessary, and indeed two of the UKWAC partners are involved in whole domain archiving: the National Archives have been working with the Internet Archive since 2003 on '.gov.uk' sites to build the UK Government web archive¹⁴; and the BL is exploring a similar approach to providing '.uk' snapshots.

LEGAL ISSUES

As a website is in effect a publication, and to archive a website is to create (at least) one complete copy of the publication, this falls within international copyright law. With only a small number of countries having passed legislation permitting blanket web archiving – see New Zealand as an example¹⁵ – whole domain archiving in particular could be challenged on legal grounds. In the UK, some organisations, such as the National Archives and the National Archives of Scotland have a legal mandate to archive government websites, but in most cases organisations must obtain formal permission to archive, and in some cases third party copyright makes obtaining permission more complex. In addition to archiving the site, permission must also be obtained to provide public access to the archived copy.

LEGAL DEPOSIT LIBRARIES ACT 2003

Although the UK has as yet no legislation permitting formal archiving of websites, the Legal Deposit Libraries Act 2003 lays the foundations for legal deposit of electronic publications, including websites.¹⁶ Indeed, the Act specifically allows for exemption for both copyright and database rights in future regulations, and this may allow for more effective archiving, with safeguards being put in place for commercial and restricted web publications. The UKWAC project provides a valuable test bed for some of the issues future regulations will need to take into consideration.

PERMISSION-BASED ARCHIVING

In the meantime UKWAC is adopting a formal permission-based approach to web archiving. This is in part due to the legal issues outlined above, but is also to ensure that, at this early stage of web archiving in the UK, we build up trust and dialogue with the web publishing community, which is made up of wide and diverse interests.

OTHER ISSUES SURROUNDING SELECTIVE, PERMISSION-BASED WEB ARCHIVING

Other issues include:

- labour intensiveness, involving selection, evaluation, cataloguing, permissions, harvesting, and maintenance of the public archive
- illegal content, e.g. libellous or inflammatory content
- requirement for interoperable descriptive, technical and preservation metadata
- limitations of current (and future) web archiving technology
- long-term preservation of digital objects
- territoriality ('place' of publication/jurisdiction)
- frequency of harvesting.

It is the aim of the pilot to identify and evaluate further issues, and to work with colleagues in the UK and around the world to identify the most appropriate solutions and strategies for dealing with them.

CURRENT PROGRESS

Archiving began in earnest at the very end of the 2004, with each partner archiving sites within their respective collection policies. The following

are just a couple of the areas that the consortium has focused on so far:

- The Tsunami Disaster ¹⁷ One of the first pilots within a pilot that the consortium tackled was to focus on the tsunami crisis in December, and to collect sites covering the disaster. Despite the unfortunate circumstances, this provided the consortium with a valuable test case of how we could tackle significant, but unexpected events, and how methodology could be adapted to deal with circumstances.
- UK General Election 2005 Party, constituency, candidate and other election-related websites are being selected as a number of the partners focus on this year's general election, building a picture of how the election developed on the web, before, during and after the 5 May.

PUBLIC ACCESS TO THE UK WEB ARCHIVE – TODAY!

By the time you read this article, the fledgling UK web archive will have gone live. ¹⁸

Although still in its infancy – only 5 months old – and with the size of the archive still relatively small, this will be the first chance that users, and the archived publishers themselves, will have to search for and view the websites archived by the consortium, and see how the archive will work in practice.

Fig 1. Home page of the UK web archive



The public interface of the archive (see fig 1 above) provides simple keyword searching and alphabetic browsing, supplemented by straight-

forward subject navigation, and the ability to access sites via collections, such as the tsuanmi disaster, which helps to bring together otherwise unrelated sites.

To supplement resource discovery via the web archive, the majority of partners are also providing catalogue records within their information management systems. This helps to integrate the archived sites into the traditional collection, as well as providing potential for greater interoperability and openness of the metadata.

THE FUTURE

UKWAC is looking at web archiving very much into the long term, and the partners are committed to carrying the work of the project forward post-pilot. In terms of technological developments, UKWAC has been working with the International Internet Preservation Consortium (IIPC) ¹⁹, who are developing a suite of tools for web archiving and digital preservation. Part of this work includes developing a possible successor to PANDAS, and this should bring benefits to the web archiving community as a whole.

Other bodies around the world are carrying out interesting research into the different areas that affect web archiving and digital preservation. For a taste of the many issues not covered in this short article, the website of the first International Web Archiving Conference – attended by a number of the UKWAC partners, including NLS – is a good starting point ²⁰.

As both UKWAC and IIPC demonstrate, collaboration will be key to success in this area, as in so many others. As such, the NLS is keen to work with partners within Scotland, and we have been in discussions with a number of key players, such as the National Archives of Scotland, to ensure that important parts of our national heritage are not lost forever. Collaboration with web publishers and authors is also key to ensuring such an important enterprise is a success, and we must not over-

look the intended users of the archive. Whether they be mainstream commercial publisher, or small website owner, professional researcher, or

curious public, their views, needs and wants must sought and listened to, and we invite interested parties to get in touch with either ourselves, or a relevant member of the consortium.

In the meantime the National Library of Scotland, with the other members of UKWAC, will continue to build a UK web archive for users of today, and tomorrow, and consult and work with colleagues, publishers, and the public to make the UK Web Archive a success.

From 9 May 2005 you can view the UK Web Archive at: <http://www.webarchive.org.uk/>. The project website is available at: <http://info.webarchive.org.uk/>.

References

- 1 Deborah Woodyard, Domain UK: Britain on the web. Power point presentation, given at '2nd ECDL : workshop on web archiving, 2002', available at: <http://bibnum.bnf.fr/ecdl/2002/uk/uk.html> (Accessed 8 April 2005)
- 2 The Internet Archive's waybackmachine, available at: <http://www.archive.org> (Accessed 8 April 2005)
- 3 National Library of Australia's Pandora website, available at: <http://pandora.nla.gov.au/> (see also reference 11) (Accessed 8 April 2005)
- 4 Nordic web archive, available at: <http://nwa.nb.no/> (Accessed 8 April 2005)
- 5 See insight into Royal Library of Sweden web archive in Kulturarw3 : long time preservation of electronic documents, available at: <http://www.kb.se/kw3/ENG/> (Accessed 8 April 2005)
- 6 See the United States Library of Congress Minerva project, available at: <http://www.loc.gov/minerva/> (Accessed 8 April 2005)
- 7 UK Web Archiving Consortium, The wonders of the web captured for ever, (UKWAC press release), available at: <http://info.webarchive.org.uk/pressrelease21-06-04.html> (Accessed 8 April 2005)
- 8 Scholars note 'decay' of citations to online references IN Chronicle of higher education, 14 March 2005, available at: <http://chronicle.com/prm/daily/2005/03/20050> (subscription required) (see also Diomidis Spinellis, The decay and failures of web references, IN Communications of the ACM, 46(1):71-77, January 2003. Draft available at: <http://www.dmst.aueb.gr/dds/pubs/jrnl/2003-CACM-URLcite/html/urlcite.html>) (Both accessed 8 April 2005)
- 9 Michael Day, Collecting and preserving the World Wide Web: a feasibility study for the JISC and the Wellcome Trust, Bath : UKOLN for Wellcome Trust and JISC, 2003, available at: <http://library.wellcome.ac.uk/assets/WTL039229.pdf> (Accessed 8 April 2005)
- 10 For an example of an archived site, see Transport Edinburgh, available at: <http://www.webarchive.org.uk/tep/10802.html>
- 11 Paul Koerbin, The Pandora digital archiving system (PANDAS) : managing web archiving in Australia : a case study, gives a very good outline of the PANDAS system, available at: <http://www.nla.gov.au/nla/staffpaper/2004/koerbin2.html> (Accessed 8 April 2005)
- 12 See <http://www.httrack.com/> for information on HTTrack. (Accessed 8 April 2005)
- 13 See Magus Research web site, available at: <http://www.magus-research.com/> (Accessed 8 April 2005)
- 14 The National Archives, UK Government web archive, available at: <http://www.nationalarchives.gov.uk/preservation/webarchive/> (Accessed 8 April 2005)
- 15 The National Library of New Zealand (Te Puna Maturanga o Aotearoa) Act 2003, available at: <http://www.natlib.govt.nz/files/Act03-19.pdf> (Accessed 8 April 2005)
- 16 Legal Deposit Libraries Act 2003, Norwich: Stationery Office, 2003, available at: <http://www.legislation.hmso.gov.uk/acts/acts2003/20030028.htm> (Accessed 8 April 2005)
- 17 See the Tsunami disaster : Indian Ocean 2005 collection on the web archive, available at: <http://www.webarchive.org.uk/col/c8050.html>

- 18 The UKWAC web archive is available from 9 May 2005, at: <http://www.webarchive.org.uk/>
- 19 International Internet Preservation Consortium website, available at: <http://netpreserve.org/about/index.php> (Accessed 8 April 2005)
- 20 National Library of Australia, Archiving web resources : issues for cultural heritage institutions : international conference, conference web site available at: <http://www.nla.gov.au/webarchiving/> (Accessed 8 April 2005)