It seems fair to say that the lion's share of work on developing online tools for reference and citation management by students and researchers has focused on familiar types of publication. They generally comprise resources that can be neatly and discretely bound in the covers of a book or journal, or their electronic analogues, like the Portable Document Format (PDF): objects in established library or database systems, with ISBNs and ISSNs underwritten by the authority of formal publication and legal deposit.

Yet, increasingly, native Web resources are also becoming eminently citable, and managing both the resources, and references to them, is an ongoing challenge. Moreover, the issues associated with referencing this kind of material have received comparatively little attention, beyond introducing the convention that includes the URL and the date it was accessed in bibliographies. While it may be hard to quantify the "average lifespan of a web page" [1], what is undeniable is that Web resources are highly volatile and prone to deletion or amendment without warning.

Web Preservation is one field of endeavour which attempts to counter the Web's transient tendency, and a variety of approaches continue to be explored. The aim of this article is to convey the fairly simple message that many themes and concerns of Web preservation are equally relevant in the quest for effective reference management in academic research, particularly given the rate at which our dependence on Web-delivered resources is growing.

Digital preservation is, naturally, a strong theme in the work of the University of London Computer Centre (ULCC)'s Digital Archives Department, and Web preservation has featured particularly strongly in recent years. This article will draw upon several initiatives with which we have been involved recently. These include: the 2008 JISC Preservation of Web Resources Project (JISC-PoWR) [2], on which we worked with Brian Kelly and Marieke Guy of UKOLN; our work for the UK Web Archiving Consortium; and the ongoing JISC ArchivePress Project [3] (itself, in many ways, a sequel to JISC-PoWR).

Another perspective that I bring is as a part-time student myself, on the MSc E-Learning programme at Edinburgh University. As a consequence I have papers to read, and write, and a dissertation imminent. So for this reason too I have a stake in making it easier to keep track of information for reading lists, footnotes and bibliographies, whether with desktop tools or Web-based tools, or through features in online VLEs, databases and repositories.

Why Do We Cite?

Why do we cite at all? This statement from Edinburgh University Information Services clearly expresses the standard explanation:

1. To allow those reading the record of what you've done to read the sources you have read.
2. To credit, and show you have read, the key relevant work and can use it to support your arguments and so indicate where your work has taken you further.
3. Citing and referencing work you have read and used avoids any charge of plagiarism. [4]

This would seem to be the most basic, essential description necessary for undergraduate or post-graduate students. It is a resolutely traditional and non-technical explanation: it does not mention advanced topics, like reverse citation tracking or real-time impact metrics for the Research Excellence Framework (REF)!

Crediting sources and avoiding plagiarism are important scholarly aims. However it is the first aim that is particularly significant as it clearly suggests that authors, researchers and scholars must try to give their readers at least a fighting chance of being able to access the same sources they consulted in arriving at their conclusions. The onus is on an author to cite in a way that
allows other people to check. A standard set of metadata - author, title, publisher, data, journal title - has evolved to provide a reasonably reliable system of 'indirection'. What is not suggested is that authors are responsible for making those referenced works available to their readers. And therefore, I think it is generally accepted, the reference should be to some authoritative edition of a work, authority attested by publication and legal deposit, or some other dependable chain-of-custody.

Web Preservation

A standard definition of digital preservation is:

\[ \text{'the series of managed activities necessary to ensure continued access to digital materials for as long as necessary.' [5]} \]

As we discovered on the JISC-PoWR Project, Web preservation is all this, with the added expectation of some kind of continuity of access to Web resources over the Web, with a reference/locator/identifier that will persist. This brings with it a number of requirements to deal with the problems of transient locators and changing identifiers which are also a feature of the Web. Unfortunately, the same features which made the Web so easy to adopt make it arguably too easy to adapt. Increasingly, in a post-Web 2.0 world, we also have highly volatile content, easily modified or removed by its authors and editors, without any guarantee that previously published versions, or any record of the change will persist.

Approaches to dealing with this range from Tim Berners-Lee's thoughts on Cool URLs, [6] to a variety of systems of Digital Object Identifiers (DOIs) and Handles and Persistent URLs. Other ideas for time-based and version-based extensions to HTTP are emerging. Many wikis, notably Wikipedia/MediaWiki, implement a versioning system, based on the Unix diff utility. Some blogging platforms, such as WordPress, support access to superseded versions of posts and pages.

At the National Archives, the Web Continuity Project [7] arose from a request by Jack Straw, as leader of the House of Commons in 2007, that government departments ensure continued access to online documents, after a survey revealed that 60% of links in Hansard to UK government Web sites for the period 1997 to 2006 were broken.

Web archiving is also an important activity in this respect. Among the most prominent endeavours in this area are the Internet Archive's Wayback Machine [8], and in the UK we have the UK Web Archive [9], begun as a joint project of the British Library (BL), The National Archives (TNA), the Joint Information Systems Committee (JISC) and the Wellcome Trust. Content in these archives is copied from its original location into the archive, with the aim that it look or function more or less as it did in its original form. We assume that, once in Web archives, Web resources will have a degree of permanence and fixity, and be reliably accessible and referenceable for a long time. But, ideally, we should not have to depend solely on the likes of BL and TNA: sustainable, distributed collections of Web resources should be as widely achievable as traditional archives. Unfortunately, at present, the various applications and components that are used to produce collections like Wayback and the UK Web Archive are not for the faint-hearted, and involve considerable investment in specialist skills and resources.

Citing Web Archives

A question I would ask at this point is: are we educating students and researchers to use these kinds of reliable collections when citing Web material? Does Information Literacy/Digital Skills training do enough to raise awareness of the issues, and the solutions?

Issues of security and trust are at the heart of information skills training requirements. In Higher Education and research there is work going on at all levels to ensure all stakeholders are aware of the importance of using trusted resources on the Web. One example is Intute's Internet Detective [10], designed for students in Further and Higher Education. (The need for trusted management of resources, in this case research outputs, is also central to the debate over Open Access.) Not everything anyone finds useful to reference will be in a trusted archive, but over time we can expect the proportion to grow. It is important to educate authors referencing Web content to locate and cite trusted versions whenever possible. And the collections should be designed to support this aim too.

By way of a personal example: in a paper I wrote last year for my MSc course I decided to cite an apposite post by Brian Kelly on the JISC-PoWR project blog. I might have cited the URL of the blog 'in the wild':


Yet such a citation is of little use if Brian decides to delete or change his post, or if anything else happens to compromise the location of the blog or the service from which it is available. Instead, in an attempt to put into practice what I preach, I cited the copy in the UK Web Archive:


I think it is the right thing to do, as this should be a permanent and fixed representation of the resource I consulted. But what a mouthful that URL is! The people at TinyURL.com know that you can represent over 2 billion objects with just 6 alphanumeric
characters, and I am confident there are not yet two billion objects in the UK Web Archive.

Reservations about the URL format aside: are other students and researchers doing this? Could they? Should they? Perhaps I only thought to do this because I had just worked on that very same project, with my colleague Ed Pinsent, who also works on the UK Web Archive Project. And it was a remark of Ed's that alerted me to copies of the JISC-PoWR blog in the UK Web Archive. I hope that it is not necessary for students to hob-nob with the inner circles of the Information Environment programme to know about such things.

The Significance of Blogs

You will notice it was a blog post that I cited, and that is significant because blogs are a particularly interesting class of Web resource, that use a common approach to achieve a format of compelling utility in a wide range of journal-like activities. For the most part they manifest themselves as public-private diaries (I'll avoid the term 'journals') and many are eminently citable. You do not have to take my word for it, this is what Peter Murray Rust has said:

Blogs are evolving and being used for many valuable activities (here we highlight scholarship). Some bloggers spend hours or more on a post. Bill Hooker has an incredible set of statistics about the cost of Open Access and Toll Access publications, page charges, etc. Normally that would get published in a journal no-one reads [...] So I tend to work out my half-baked ideas in public. [11]

Michael Nielsen is another noted commentator on scientific scholarship:

It's easy to miss the impact of blogs on research, because most science blogs focus on outreach. But more and more blogs contain high quality research content. [...] What we're seeing here is a spectacular expansion in the range of the blog medium. By comparison, the journals are standing still. [12]

With this affirmation of the importance of blogs in mind, and the work on JISC-PoWR, and my own studies, I was also struck by a recent article of Heather Morrison's in First Monday ("Rethinking Collections" [13]). Heather described blogging as representing a new communications paradigm, and pointed out that it was the blog format - not the academic journal - that made an important endeavour like Peter Suber's Open Access News [14] possible and useful. Heather also made the point that institutional libraries should look beyond their traditional approaches to collections:

The discrete "item" — the book, the journal article — is becoming less and less relevant in today's interconnected world. [...] For the library, what this means is that collections work will gradually need to shift from a focus on discrete items, to a focus on comprehensive collections and links both within and outside of collections. [13]

The idea of creating and maintaining collections of researchers' blogs seems a sensible and logical one, particularly within the context of the institutional record and remit. Academic institutions, after all, have the degree of authority and longevity appropriate to managing such a collection in the medium-to-long term. After all, they generally manage archives already.

One advantage of the institutional context is that it ought considerably to reduce the rights-related issues surrounding the keeping of copies of Web material. Dealing with copyright issues is a considerable overhead, but within the context of research done under the aegis of a particular institution, it ought to be relatively easy to establish that the institution has rights in material created by its employees or affiliates, or on its servers. This constraint has already been exploited quite effectively by Institutional Repositories and Open Access endeavours.

The ArchivePress Project

This was also, in part, the inspiration for the ArchivePress Project, part of the JISC Rapid Innovation programme, which we have been developing at ULCC, with the invaluable support of Maureen Pennock, of the British Library's digital preservation team. Its simple premise is to develop and testbed an easy way for any institution, group or individual to create a reliable working collection of posts on multiple blogs. ArchivePress uses the WordPress blogging system, and is building on existing third-party plug-ins to harvest posts from RSS and Atom news feeds - which are standard across all blogging platforms.

This does not address or solve every issue of Web or blog preservation. It focuses predominantly on the text content: it can still preserve many hypertextual features, but not the look and feel, nor the bells and whistles, of any specific blog platform, theme or instance.

One justification for this approach is simply that it is easier to do this way, compared with the industrial strength harvesters of the big Web archives. It was a conversation with Brian Kelly, when setting up the JISC-PoWR blog, that drew my attention to the fact that many 'power-users' use aggregating feed-reader programs to access blog content. Chris Rusbridge (Director of the Digital Curation Centre) also expressed sympathy with this approach:

[Design may be an issue; on the other hand I rarely see any blog's design, since I read through NetNewsWire, so I'm inclined to think blogs represent an area where the content is primary and design secondary. [15]

Longer-term preservation of the archive created is an issue we are sidestepping, but establishing an effective framework for creating collections in an automated, straightforward way is a necessary first step. All the technology involved is open source...
and uses open Web standards for content and metadata, so in that respect we can be fairly confident that collections so created are receiving a sustainable start. We expect working plug-ins to be available early in 2010.

There are, no doubt, other and better ways to preserve institutional Web content. Increasing use of multiple-blog systems, like the WordPress Multi-User platform, and improvements to Content Management Systems, offer an opportunity to build more sustainability features into new systems being rolled out. But I do not think Institutions should wait too long before they start capturing valuable blog content as part of the institutional record; moreover, ideally, they should also be promoting the use of those archived collections.

Citation Archives

A related Web preservation endeavour is the Citation Archive. At the Missing Links workshop at the British Library in July 2009 [16], we heard a fascinating account by Hanno Lecher (Librarian at the Sinological Library, Leiden University) of the DACHS Citation Archive at the Chinese Studies departments of the Universities of Leiden [17] and Heidelberg [18]. Driven by the particular volatility of the Chinese Web, these Chinese studies departments have created a system that will keep snapshots of materials that their researchers reference. The aim is to capture and archive relevant resources as primary source for later research.

The repository stores a copy of a requested page, or pages, along with appropriate metadata (including its original URL). Not surprisingly, many objects stored in the repository are no longer available at their original URL: pages on the Chinese Web are often literally here today, gone tomorrow. Because of this, DACHS has been developed along with strict protocols for verifying URL references, evaluating reliability of online resources, and forbidding references to unsecured online resources outside the archive. Although this kind of activity can, theoretically, lead to copyright issues, the principle of caching Web content appears to be well-established, not least by Google and the Internet Archive: an accessible take-down or opt-out policy (in the event that a copyright owner objects to the archive making a resource publicly available) is usually considered sufficient safeguard.

One public service that offers these features is WebCite, hosted at the University of Toronto [19]. It has been adopted as a standard for Web references by a number of journals, notably those published by BioMed Central. It is available for general use and offers automated features that facilitate robust citations including both the original URL and the URL of the archived copy. A simple transaction with the WebCite submission form (or browser bookmarklet) yields the kind of reference for which I think we should be striving, for example:


Conclusions

As ever more information resources become available on the Web, the need for effective preservation solutions continues to grow. The case therefore also grows for academic authors - students and researchers - to acquire the habit of referencing stable, reliable copies of them, rather than copies ‘in the wild’, which can easily mutate, or disappear without trace. There may be no onus on authors to manage copies of the resources they reference, but they should aim to cite objects with a reasonable degree of authority and permanence.

There is therefore also a strong case to be made for accessible Web archives that address the needs of academic authors. Existing Web archives should ensure the student/researcher use case is adequately reflected in their system specification and provide robust functions and features appropriate to their use as scholarly resources. Persistent and succinct identifiers/locators, for one thing, as well as embedded rich metadata, would ensure reference management tools can work as effectively with Web archives as they do with established sources of literature.

In line with the conclusions of the JISC-PoWR Project, we should continue to encourage institutional implementation of Web applications supporting persistence, versioning and sustainability - including Content Management Systems, blogs and wikis. Institutional Web archive collections should also be encouraged, developing the new kinds of electronic collection suggested by Heather Morrison and implicit in recent initiatives like the Libraries of the Future campaign [20]. A critical mass of collections like this might in turn enable richer and even more advanced applications to become mainstream; text-mining, for example, or hypertextual content transclusion, as recently described by Tony Hirst [21].

Most of all we need to ensure that information skills and literacy efforts spread the message about the importance of Web resources - and the risks associated with them - and ensure that current and future generations of students and researchers have the knowledge and tools necessary to identify and cite reliable, authentic, persistent Web resources wherever possible.

References

2. JISC-PoWR (Preservation of Web Resources): a JISC-sponsored project http://jiscpowr.jiscinvolve.org/
3. ArchivePress: a JISC-sponsored project http://archivepress.ulcc.ac.uk/
4. Edinburgh University MSc E-learning Course Handbook 2006
6. Tim Berners-Lee: 'Cool URIs don't change' http://www.w3.org/Provider/Style/URI
7. The National Archives Web Continuity Project
   http://www.nationalarchives.gov.uk/webcontinuity/
9. UK Web Archive http://www.webarchive.org.uk/ukwa/
10. Intute: Internet Detective http://www.vts.intute.ac.uk/detective/
11. Peter Murray Rust blog post: Effective digital preservation is (almost) impossible; so Disseminate instead; comment by Richard Davis, 26 June 2009
   http://wwmm.ch.cam.ac.uk/blogs/murrayrust/?p=2159
12. Michael Nielsen blog post: Is scientific publishing about to be disrupted? 29 June 2009
   http://michaelnielsen.org/blog/is-scientific-publishing-about-to-be-disrupted/
13. Heather Morrison, Rethinking collections - Libraries and librarians in an open age: A theoretical view,
    First Monday, Volume 12 Number 10 - 1 October 2007
15. Comment by Chris Rusbridge to post by Gavin Baker entitled Preservation for scholarly blogs, 30 March 2009
   http://www.gavinbaker.com/2009/03/30/preservation-for-scholarly-blogs/
17. DACHS at Leiden University http://leiden.dachs-archive.org/citrep/
18. DACHS at Heidelberg University http://www.sino.uni-heidelberg.de/dachs/
20. JISC Libraries of the Future campaign http://www.jisc.ac.uk/librariesofthefuture
21. Tony Hirst blog post, Content transclusion one step closer

Author Details

Richard M. Davis
Repository Service & Development Manager
Digital Archives Department
University of London Computer Centre (ULCC)

Email: r.davis@ulcc.ac.uk
Web site: http://dablog.ulcc.ac.uk/author/richarddavis/

Return to top

Article Title: "Moving Targets: Web Preservation and Reference Management"
Author: Richard M. Davis
Publication Date: 30-January-2010 Publication: Ariadne Issue 62
Originating URL: http://www.ariadne.ac.uk/issue62/davis/

This article has been published under copyright; please see our access terms and copyright guidance regarding use of content from this article. See also our explanations of how to cite Ariadne articles for examples of bibliographic format.

Ariadne is published by UKOLN. UKOLN receives support from JISC and the University of Bath where it is based.
© Ariadne ISSN: 1361-3200. See our explanations of Access Terms and Copyright and Privacy Statement.

Source URL (retrieved on 26 May 2012 - 11:06): http://www.ariadne.ac.uk/issue62/davis