Internet Archiving – The Wayback Machine

Ludovica Price

School of Informatics

City University London

Introduction

In a world where information is just a mouse click away, the concept of information as resource is more relevant today than it has ever been before. Information has a price as much as water, gas and electricity does. As Floridi notes, "[n]obody pays for yesterday's newspaper or the wrong kind of information… Information has economic value because of its usefulness" (2010, p.90). But is information a resource that is *managed* in the same way as, say, gas or water?

This is a broad question and cannot easily be answered in a short space. As such, this essay will focus on information as an *internet* resource; specifically on the archiving of web pages via the Internet Archive.[1] It will explore the issues unique to web archiving, particularly the ephemeral nature of internet information resources, the legal issues pertaining to them, and the practical problems involved in the Internet Archive's selection and collection methods. The aim of this essay is to demonstrate that information as a resource – particularly pertaining to the treatment of internet resources – faces unique problems that cannot easily be mapped onto the treatment of other resources.

What is the Internet Archive?

The Internet Archive is essentially an internet library, begun in 1996 by Brewster Kahle and Bruce Gilliat, and created with the goal to archive the entire World Wide Web for the use of historians, scholars and researchers. Their goal is to provide "universal access to all knowledge" through the preservation of "Web pages, books, music, software, and moving images" (Kimpton and Ubois, 2006, p.201).[2] Since 2001 it has granted public access to its databases through its search interface, the

[1] <http://www.archive.org/web/web.php>.

[2] A broader overview of the Internet Archives mission may be viewed on their website at <http://www.archive.org/about/about.php>.

Wayback Machine. Though non-profit, it gathers its data using software donated by Alexa Internet and others. According to the site,[3] the archive has collected 2 petabytes worth of data and is growing at the rate of 20 petabytes per month.

The Archive explicitly states that part of their mission is to "prevent the internet… and other 'digital-born' materials from disappearing into the past", to turn the internet from "ephemera into artefact" (Internet Archive, 2011). This in many ways is the crux of the Internet Archive; yet it is a tall order. The Archive was essentially the first to rise to the challenge of archiving the internet in its entirety, and while others have since followed in their footsteps, the internet's rate of growth (and disappearance) makes it a formidable task – perhaps an impossible one. So how does the Internet Archive approach this unique challenge, and with what success? And what problems has it faced in its pioneering attempt to turn 'ephemera into artefact'?

The Internet as ephemera

In a 2000 study by Alexa Internet (cited in Day, 2003), it was estimated that the average lifespan of a web page was between 75 and 100 days. Online information is easily deleted, overwritten and made redundant. Immediacy of access is of greater importance than permanence. What this means is that large swathes of information remain unpreserved and are quickly lost under the tidal wave of mass deletion, editing and so forth.

Already much of the internet's content was lost when the Internet Archive began archiving it in 1996. But it is the constantly shifting nature of the digital landscape that makes it so difficult to pin down and preserve. How do you preserve the tides of an ever-changing sea?

Methods of collection

There are different methods of preserving or 'collecting' the internet. These are direct transfer, remote harvesting, database archiving and transactional archiving (Brown, 2006). Each has its own strengths and weaknesses, but we shall choose to focus on remote harvesting, since this is the method used by the Internet Archive. [4] Indeed, it is the most widely used method of collecting websites.

---

[3] See <http://www.archive.org/about/faqs.php#9>.

[4] Adrian Brown gives an in-depth description of each collection method in his book, *Archiving Websites* (2006) by Facet Publishing.

Remote harvesting is accomplished through web crawlers – 'robots' or 'spiders' – the same technology which is used to index the internet by search engines. Intervention is minimal in that the web crawler is able to automatically retrieve content with or without the input of set parameters. Collection is based on a path-finding system. This means that the robot will scan the web page, identify hyperlinks, and collect all the linked pages – and so on. The Internet Archive uses the Heritrix web crawler software, which was specifically created by the Internet Archive with partner institutions (Rackley, 2009).

The advantage of remote harvesting is that it does not discriminate in terms of subject matter (unless set parameters command it to), and it is also a fast method of data collection. In this sense, the Internet Archive may be seen to be a 'true' archive, not merely a group of 'special collections.' In its early days, the Archives was able to capture the entire net (Lessig, 2004, p.108). As one 'round' of the internet was completed, another would immediately follow. But the size of the internet, its constant growth and mutability has since made net-wide crawls virtually impossible. It is inevitable that not all changes to a website will be captured. Like a water leak, some information will slip through the fingers of the web crawlers, and will be subsequently lost forever. Therefore the Archive cannot be considered to be 'complete'.

Another practical problem with remote harvesting is quality. Crawlers are limited in what they can capture, and what they capture is essentially a static snapshot of a site. There is not yet adequate technology to capture dynamic content. Java or Flash-based functions will not work; neither will search facilities or logon scripts. This means that a great deal of functionality is lost from the original webpage. This can compound the problems posed by only partial captures of websites. As Day (2003) found in his study of the Archive for the Wellcome Trust, some archived sites are incredibly limited, almost un-navigable.

This begs the question – how much of the information stored in the Archive is actually useable, let alone useful?

The 'Deep Web'

An enormous proportion of the web is supposedly made up of the 'deep web', that is, sites that are 'invisible' or 'hidden' to robots. These sites may be hidden by password protection, require a certain query to be retrieved, or be contained in a database (Brown, 2006). In 2001, Bergman estimated that

the deep web was 550 times the size of the 'surface' web (cited in Masanès, 2006). What proportion it takes up now is largely unknown.

Due to the nature of its collection methods, the Internet Archive is unable to capture this information – information that is arguably richer than that which is contained by the 'surface' web. At any rate, a vast proportion of the net remains unarchived, and whilst studies into the possibility of archiving the hidden web have been looked into (Masanès, 2006), as yet no attempt has been forthcoming. For a site whose vision is to archive the entire net, this is a serious shortcoming, and emphasises just how much valuable information is lost before it is even found – a resource that may well have never even existed.

Presentation and access

The Internet Archive can be accessed by the public via its search interface, the Wayback Machine, which was created by Alexa Internet in 2001. The Wayback Machine requires the URL of the website in question to be entered into its search engine; its search results are time-based, showing archived snapshots of the web page in chronological order (see Figure 1). The user may surf the site in its static, archived form; the application edits links to prevent redirection to the 'live' web.

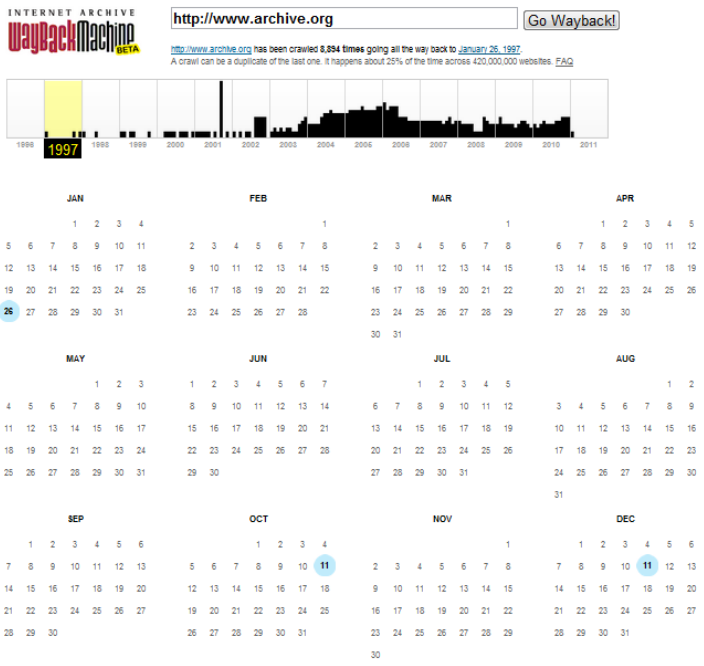

Figure 1 - The current interface for the Wayback Machine, accessed on 3 January 2012. Archived search results for the Internet Archive's website are shown through a time graph and calendar. As can be seen, results from 2011 are mostly unavailable.

However, some access problems remain – for example, the URL of the site in question must be known, as there is no way of searching by title. Another problem is that since the remote harvesting of websites is based on path-finding (via hyperlinks), there is no one complete snapshot of a website taken at a single point in time (Rackley, 2009). For instance, while navigating a page captured on 1 January 2001, one may click on a link that was not archived at the time, but at a later date, say, 1 January 2002. The user will be taken to a page that contains information from a later date and is thus anachronistic. Therefore it is possible that a certain loss of context is inevitable when navigating some sites.

Preservation

Digital material may be copied perfectly an infinite amount of times, and back up files are a regular part of digital life. This would make it seem that the preservation of digital data is easy. In fact, this is far from being the case.

Digital preservation poses a unique challenge. As Charlesworth notes: "The media on which digital material is held may also be subject to deterioration over time, requiring the transfer of data to new storage media and possibly new formats" (2003, p.9-10). Since digital data is easily manipulated, it is also easily corrupted, and maintaining integrity is a problem. A greater problem is that technology is always evolving, with older technologies constantly becoming obsolete[5]. Once a website is archived, how is it possible to keep it accessible? Brewster Kahle acknowledges this problem himself:

> "The current digital technologies only last about three years. In the last ten years, we've moved – transitioned – our materials three times. It's painful. And lossy. It's very difficult." (2007, p.30)

Migration and emulation are standard preservation techniques – migration involves the conversion of files to formats supported by existing technologies, and emulation involves the recreation of obsolete technologies on existing computer platforms (Brown, 2006). Both these methods involve extensive technological skill and effort. Having to perform them every couple of years is an ongoing process, and there is currently no quick and easy way of preserving digital content.

[5] Moore's Law states that computer processing power has doubled every 18 months since the 1970s (Brown, 2006).

Currently, the Internet Archive makes use of both migration and emulation; it also attempts to avoid disaster or accidents through the maintenance of mirror sites[6]. It seems that the only way in which the preservation problem can be fully resolved is when more stable, long-term technologies can be developed to aid the web archivist in their task (Kahle, 2007, p.30).

Copyright and other legal issues

The Internet Archive aims to archive the internet in its entirety. But, as Alan Yentob notes in his TV programme *Imagine: Books – The Last Chapter?*, this is "subject, of course, to the objection of copyright holders" (BBC, 2011).

Of course, technically, everyone is a copyright holder (Charlesworth, 2003). This extends to material stored on websites, and thus "downloading and storage would inevitably involve the creation of unlicensed copies of the work that went to make up the webpage" (Charlesworth, 2003, p.7). The commonsense way of getting around this is to ask the creator of the website permission to copy their work. Of course, this is an impossible task for an information resource as vast as the web. When such information is being copied via automatic remote harvesting it is clearly not a viable option.

The Internet Archive currently follows the Oakland Archive Policy (Kimpton and Ubois, 2006)[7], which recommends the removal or access restriction of archived material upon request. The Internet Archive accomplishes this in two ways. The first is by giving web-owners the choice to 'opt-out' via the standard robots.txt, which excludes websites from remote harvesting by robots[8]. The second is by giving rightholders the option to request that their sites be withdrawn by the Archive.

The fact that the Internet Archive does not request permission to copy websites *before* remote harvesting puts it at considerable legal risk. Most rightholders may be satisfied with an a-priori or a-posteori opt-out system, but assuming that *all* will be satisfied is not necessarily a safe way to operate. Neither system overrides the rightholder's automatic copyright, and offering to take down copied material after the fact is not recognised as sufficient protection should a law suit be brought forward (Charlesworth, 2003).

To date, lawsuits have been brought against both the Internet Archive and Google's similar web page archive, the Google cache. The Google cache is not an archive in the true sense – it is more of a

[6] See <http://www.archive.org/about/about.php#preserve> for more information on the Internet Archive's preservation techniques.
[7] The policy is outlined at <http://www2.sims.berkeley.edu/research/conferences/aps/removal-policy.html>.
[8] See <http://www.robotstxt.org/robotstxt.html>.

back-up to be used when a page is offline, deleted, or links are broken. It is refreshed every fourteen to twenty days (Copyright Website, 2006). Google was sued in 2006 by Blake Field, a lawyer and author who objected to Google caching his poems on the basis that doing so was a copyright infringement. The case was subsequently thrown out of court by the judge, who concluded that Field had known how to disable the cache of his work yet had chosen not to do so. More importantly, the judge declared "Google's use of copyrighted material was a fair use" (Pinsent Masons, 2006).

Ironically, a suit was brought against the Internet Archive because it was used in another suit (Zellmer Jr., 2005)[9]. In 2003, law firm Earley Follmer defended health insurance company Health Advocate in a trademark action against a company using a similar name, Healthcare Advocates. Earley Follmer used the Wayback Machine to prepare their case; two years later, Healthcare Advocates sued both Earley Follmer and the Internet Archive on the grounds of copyright infringement. Since the plaintiff had used robots.txt to block crawling of its website, the Archive had technically breached its tacit agreement with the company. Ultimately, the case was settled out of court.

Legally, it would seem that digital archives are caught between a rock and a hard place. As the law stands, it is not entirely clear what a web archive's limitations are, nor what its legal defence might be. To some extent at least, the rights of web archivists are protected by legal deposit laws. In the US, offline publications are subject to legal deposit in a static, physical form (Charlesworth, 2003). In the UK, legal deposit regulations are still under review and hope to bring websites under their remit, though as yet permission is still required of rightholders to archive a website (UK Web Archive, 2011). In the meantime there is the failsafe of restricting public access to material, but this renders the archive's role *as* an archive moot.

Conclusion

It is evident from the issues discussed that information as a web resource faces many unique problems, and the Internet Archive faces unprecedented challenges in capturing and storing it. Firstly, the web is fluid - its impermanence is even acknowledged by the Internet Archive itself – mirror sites of the Archive exist in Alexandria, Egypt and Amsterdam, The Netherlands (Rackley, 2009).

Secondly, digital information and its supporting technologies face rapid obsolescence, and the threat of corruption. Web archives face an ongoing struggle in their efforts to fight the tide of

[9] It would seem that the Wayback Machine is a popular tool used by lawyers to prepare their cases; see Fagan (2007).

technological advancement. This seems to be a battle that can only be won when corporations like Microsoft turn their attention away from the marketing of commercial software to the development of stable, long-term software to aid archivists and the preservation of digital information (Kahle, 2007).

Lastly, legal frameworks must be put in place to clarify the position of intellectual property rights in the context of digital information. To date, governments have been slow to act upon the need to archive websites, and thus the process of archiving the web has been a patchy, knee-jerk reaction to 'rescue' ephemeral information, often without web archivists understanding the legal issues, or continuing to harvest data until it runs afoul of the rightholders or authorities (Charlesworth, 2003). This is a situation that must be rectified as soon as is possible.

To summarise, digital information is a resource like no other. Gas, water, electricity – though not static, they are stable, predictable resources. We know how to handle them, we know their benefits, limitations, dangers. Digital information is still a grey area, essentially an unknown. It is by no means certain just how large the deep web is, or what it contains. An edited file supersedes and replaces its predecessor leaving virtually no trace. An entire website may grow and die without anyone even knowing, rendering its contents irretrievable. The Internet Archive and other web archives have sought to solve these issues with varying degrees of success. They have pioneered a discipline almost unwittingly, rising to challenges only as they are discovered, finding solutions to problems as they arise.

Ultimately, the management of internet resources is a process which is still in its infancy, an immature discipline where the nature of the resource it is attempting to manage is only just beginning to be understood. It can be hoped that with a better comprehension of the web, its complexities and its limitations, more appropriate methods of archiving it will be forthcoming. To achieve this, support will be needed from governments, copyright laws, soft- and hardware manufacturers – and from contributions of the public itself.

Bibliography

Brown, A., 2006. *Archiving websites: a practical guide for information management professionals*. London: Facet Publishing.

Charlesworth, A., 2003. *Legal issues related to the archiving of Internet resources in the UK, EU, USA and Australia: a study undertaken for the JISC and Wellcome Trust.* [online] Bristol: University of Bristol. Available at: <http://www.jisc.ac.uk/uploaded_documents/archiving_legal.pdf> [Accessed 11 November 2011].

Copyright Website, 2006. *Field v. Google.* [online] Available at: <http://www.benedict.com/digital/internet/field/field.aspx> [Accessed: 24 October 2011].

Day, M., 2003. *Collecting and preserving the World Wide Web: a feasibility study undertaken for the JISC and Wellcome Trust*. [online] Bath: UKOLN; University of Bath. Available at: <http://library.wellcome.ac.uk/assets/wtl039229.pdf> [Accessed 11 November 2011].

Fagan, M., 2007. *"Can You Do a Wayback on That?" The Legal Community's Use of Cached Web Pages In and Out of Trial.* [pdf] Available at: <http://www.bu.edu/law/central/jd/organizations/journals/scitech/volume131/documents/Fagan_WEB.pdf> [Accessed 24 November 2011].

Floridi, L., 2010. *Information: A Very Short Introduction.* Oxford: Oxford University Press.

Imagine: Books – The Last Chapter?, 2011. [TV programme] BBC, BBC1, 13 December 2011, 22:35.

Internet Archive, 2011. *Internet Archive: Wayback Machine.* [online] Available at: <http://www.archive.org/web/web.php> [Accessed 31 December 2011].

Janes, J., 2004. Internet Librarian. *American Libraries* [e-journal] 35 (8), p.72. Available through: JSTOR. [Accessed 18 October 2011].

Jones, R., 2006. *Internet Forensics*. California: O'Reilly.

Kahle, B., 2007. Universal Access to All Knowledge. *Society of American Archivists* [e-journal] 70 (1), pp.23-31. Available through: JSTOR. [Accessed 18 October 2011].

Kimpton, M., and Ubois, J. Year-by-Year: From an Archive of the Internet to an Archive on the Internet. In: Masanès, J., ed. 2006. *Web Archiving*. New York: Springer. Ch. 9.

Lessig, L., 2004. *Free Culture: How Big Media Uses Technology and the Law to Lock Down Culture and Control Creativity.* [online] New York: The Penguin Press. Available at: <http://www.free-culture.cc/> [Accessed 25 November 2011].

Masanès, J., ed. 2006. *Web Archiving*. New York: Springer.

Pinsent Masons LLP, 2006. *Google cache does not breach copyright, says court.* [online] Available at: <http://www.out-law.com/page-6571> [Accessed 24 October 2011].

Rackley, M., (2009). Internet Archive, *Encyclopaedia of Library and Information Sciences,* 3rd ed. [online] Available at: <http://ia600606.us.archive.org/22/items/internetarchive-encyclis/EncycLisInternetArchive.pdf> [Accessed 1 January 2012]

SEOMike, 2010. When to Stop Caching. *SEOMike Search Marketing Blog,* [blog] 23 February. Available at: <http://www.customblogging.com/seomike/archives/seo-tips-tricks/when-stop-caching> [Accessed 1 January 2012].

UK Web Archive, 2011. Frequently asked questions. [online] Available at: <http://www.webarchive.org.uk/ukwa/info/faq> [Accessed 2 January 2011].

Zeller Jr., T., (2005). Keeper of Expired Web Pages Is Sued Because Archive Was Used in Another Suit. *The New York Times*, [online] 13 July. Available at: <http://www.nytimes.com/2005/07/13/technology/13suit.html> [Accessed 24 October 2011].