

# Web-Archiving

Maureen Pennock

DPC Technology Watch Report 13-01 March 2013

Series editors on behalf of the DPC  
Charles Beagrie Ltd.



Principal Investigator for the Series  
Neil Beagrie



Digital **Preservation** Coalition

DPC Technology Watch Series

**© Digital Preservation Coalition 2013 and Maureen Pennock 2013**

Published in association with Charles Beagrie Ltd.

**ISSN: 2048 7916**

**DOI: <http://dx.doi.org/10.7207/twr13-01>**

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, without the prior permission in writing from the publisher.

The moral right of the author has been asserted.

First published in Great Britain in 2013 by the Digital Preservation Coalition

## Foreword

The Digital Preservation Coalition (DPC) is an advocate and catalyst for digital preservation, ensuring our members can deliver resilient long-term access to digital content and services. It is a not-for-profit membership organization whose primary objective is to raise awareness of the importance of the preservation of digital material and the attendant strategic, cultural and technological issues. It supports its members through knowledge exchange, capacity building, assurance, advocacy and partnership. The DPC's vision is to make our digital memory accessible tomorrow.

The *DPC Technology Watch Reports* identify, delineate, monitor and address topics that have a major bearing on ensuring our collected digital memory will be available tomorrow. They provide an advanced introduction in order to support those charged with ensuring a robust digital memory, and they are of general interest to a wide and international audience with interests in computing, information management, collections management and technology. The reports are commissioned after consultation among DPC members about shared priorities and challenges; they are commissioned from experts; and they are thoroughly scrutinized by peers before being released. The authors are asked to provide reports that are informed, current, concise and balanced; that lower the barriers to participation in digital preservation; and that they are of wide utility. The reports are a distinctive and lasting contribution to the dissemination of good practice in digital preservation.

This report was written by Maureen Pennock and is published by the DPC in association with Charles Beagrie Ltd. Neil Beagrie, Director of Consultancy at Charles Beagrie Ltd, was commissioned to act as principal investigator for, and managing editor of, this series in 2011. He has been further supported by an Editorial Board drawn from DPC members and peer reviewers who comment on text prior to release: William Kilbride (Chair), Neil Beagrie (Managing Editor), Janet Delve (University of Portsmouth), Sarah Higgins (University of Aberystwyth), Tim Keefe (Trinity College Dublin), Andrew McHugh (University of Glasgow), Amanda Spencer (The National Archives) and Dave Thompson (Wellcome Library).

This report has been endorsed by the International Internet Preservation Consortium.

## Acknowledgements

I am grateful to many people for their input to this report, particularly Helen Hockx-Yu (British Library), Leila Medjkoune and Chloé Martin (IMF), and Mark Middleton (Hanzo Archives) for their assistance in compiling the case studies featured in Section 7. Thanks also go to Lewis Crawford (British Library) and others in the web archiving community with whom I have worked over the past few years and who have all helped shaped my thinking about web archiving in one way or another. Finally, the numerous projects and web archiving initiatives featured in this report must also be acknowledged, for they have played the largest role of all by bringing web archiving technology to the level of maturity we see today.

## Contents

Abstract .....	1
Executive Summary .....	2
1. Introduction.....	3
1.1. Why archive websites? .....	3
1.2. The challenge of web archiving.....	4
1.3. Key web archiving initiatives .....	5
1.4. Technical approaches to web archiving .....	6
2. Issues .....	9
2.1. Legal.....	9
2.2. Selection.....	10
2.3. Limitations of current crawlers .....	11
2.4. Authenticity, Integrity and Quality Assurance. ....	12
2.5. Archiving Web 2.0 .....	12
2.6. Temporal coherence .....	13
2.7. Viruses and malware.....	13
2.8. De-duplication .....	14
2.9. Search limitations .....	15
2.10. Long-term preservation .....	15
3. Standards .....	17
4. Software .....	19
4.1. Integrated systems.....	19
4.1.1. PANDAS.....	19

4.1.2.	Web Curator Tool (WCT) .....	20
4.1.3.	NetarchiveSuite .....	20
4.2.	Third party/commercial services .....	21
4.3.	Crawlers .....	22
4.4.	Searching .....	23
4.5.	Access .....	24
4.6.	Other options.....	25
5.	Case Studies .....	26
5.1.	The UK Web Archive.....	26
5.2.	The Internet Memory Foundation.....	28
5.3.	The Coca-Cola Web Archive.....	30
6.	Conclusions and Recommendations .....	33
7.	Glossary .....	35
8.	References and Bibliography .....	37

## Abstract

Web archiving technology enables the capture, preservation and reproduction of valuable content from the live web in an archival setting, so that it can be independently managed and preserved for future generations. This report introduces and discusses the key issues faced by organizations engaged in web archiving initiatives, whether they are contracting out to a third party service provider or managing the process in-house. It follows this with an overview of the main software applications and tools currently available. Selection and deployment of the most appropriate tools is contextual: organizations are advised to select the approach that best meets their business needs and drivers, and which they are able to support technically. Three case studies are included to illustrate the different operational contexts, drivers, and solutions that can be implemented.

This report is intended for those with an interest in, or responsibility for, setting up a web archive, particularly new practitioners or senior managers wishing to develop a holistic understanding of the issues and options available.

## Executive Summary

The World Wide Web is a unique information resource of massive scale, used globally. Much of its content will likely have value not just to the current generation but also to future generations. Yet the lasting legacy of the web is at risk, threatened in part by the very speed at which it has become a success. Content is lost at an alarming rate, risking not just our digital cultural memory but also organizational accountability. In recognition of this, a number of cultural heritage and academic institutions, non-profit organizations and private businesses have explored the issues involved and lead or contribute to development of technical solutions for web archiving. This report provides a state-of-the-art overview of the issues commonly faced and the technology in use.

Business needs and available resources are fundamental considerations when selecting appropriate web archiving tools and/or services. Other related issues must also be considered: organizations considering web archiving to meet regulatory requirements must, for example, consider associated issues such as authenticity and integrity, recordkeeping and quality assurance. All organizations will need to consider the issue of selection (i.e. which websites to archive), a seemingly straightforward task which is complicated by the complex inter-relationships shared by most websites that make it difficult to set boundaries. Other issues include managing malware, minimizing duplication of resources, temporal coherence of sites and long-term preservation or sustainability of resources. International collaboration is proving to be a game-changer in developing scalable solutions to support long-term preservation and ensure collections remain reliably accessible for future generations.

The web archiving process is not a one-off action. A suite of applications is typically deployed to support different stages of the process, though they may be integrated into a single end-to-end workflow. Much of the software is available as open source, allowing institutions free access to the source code for use and/or modification at no cost. Options range from integrated systems such as the Web Curator Tool and Netarchive Suite, to the Heritrix web crawler, the SOLR search engine, and the Wayback access interface. Other solutions are available from commercial service providers or as free small-scale online services. There are plenty of options available, but clarity over business and archiving needs is essential prior to selecting a solution to ensure these needs are met in an effective and efficient manner.

This report is aimed at curators and managers who are not technical experts but who wish to broaden their knowledge of web archiving prior to embarking on, or revising, a web archiving initiative. It will appeal mainly to organizations or individuals who are relatively new to web archiving, though existing practitioners will also find value in the summative nature of the report.

## 1. Introduction

The World Wide Web is a unique information resource. It hosts millions of websites that connect communities and individuals around the world, using cutting edge web technology that simultaneously supports its own technological advancement. The speed at which the web has become part of everyday life is unprecedented: it has taken little over two decades for the web to grow from a relatively small service used mainly by scientists, to a global information medium. It is now not only a communications hub, but also a unique record of twenty-first century life. Yet the very speed at which it develops poses a threat to our digital cultural memory, its technical legacy, evolution and our social history.

In recognition of this very real threat, organizations from around the world – particularly from the cultural heritage sector – have invested heavily in developing and implementing the technical infrastructure to support large-scale web archiving solutions. An ever-growing international web archiving community continues to actively develop new tools and techniques to improve existing capabilities, and to address the inevitable loss of access to content caused by the ephemeral nature of web content and the way it is used. This report provides an overview of that work, the state of the art in web archiving technology today, and the main issues involved.

### 1.1. Why archive websites?

Regardless of its form, there are many reasons why we archive information. Fundamental is the recognition that an information object holds value outside the original purpose for which it was created. Archives and the archiving process offer an infrastructure in which that information object can be safeguarded for as long as deemed necessary.

One of the main reasons for archiving websites, particularly in the cultural heritage community, is the relatively short timeframe within which content can be, and has been, ‘lost’: the web may be ubiquitous, but websites are transitory. Various studies have looked into the average lifespan of web pages, with results ranging from 44 days to 75 or 100 (Kahle, 1998; Lawrence *et al.*, 2001; Weiss, 2003). Specific content on a page can disappear even more frequently, particularly on news-driven and social media websites. In some cases, content can be inaccessibly ‘archived’ or relocated when sites are re-designed. In other cases, content simply disappears from the web when overwritten to ensure users have easy access to more up-to-date information. Broken links and 404 ‘page not found’ errors are the modern day equivalent of the ‘lost’ book in the library catalogue, yet far more prevalent.

Despite this, simply preserving web content ‘because otherwise it will be lost’ is a weak business case. Some institutions are legally obligated to capture and archive web content. This legal obligation is a far stronger business driver. The National Archives (UK) and the Public Record Office of Northern Ireland (PRONI), for example, are obligated by the Public Records Act (1967) to capture important records created by UK government. Government increasingly publishes its records online and the UK Government Web Archive has a very high rate of use, with over 100 million hits every



month. This is due in no small part to its use of redirection technology on live sites, pointing directly to content in the UK Government Web Archive instead of delivering 404 ‘page not found’ results (The National Archives (UK), 2010). Other legislation requires retention of records across the range of commercial and private sectors (for example the Electronic Communications Act and various pieces of discrimination legislation). Furthermore, there are clear business drivers to archive websites in the expectation of legal deposit legislation for legal deposit libraries in the UK, to ensure the preservation of UK digital cultural heritage published on the web.

Other web archiving organizations – such as the Internet Archive – collect and preserve content from the web not because of legislative requirements but because of a social interest that leads them to record the evolution and content of the Internet in its entirety and make it available to users. Use cases associated with web archives can provide other business drivers: the UK Government Web Archive, for example, is used to provide a Digital Continuity Service for government departments. Scholarly and academic research needs can also form strong use cases: academic works published online and outside the usual publications framework (such as in collaborative wikis or blogs) often fall outside a traditional collecting policy but contain work unpublished elsewhere that should arguably be preserved and used as part of the scholarly record (NDIIPP, 2012). Several studies have explored use of the Internet Archive for social science research (Dougherty *et al.*, 2010) and the scholarly value of content in web archives is increasingly recognized. Preservation of scholarly citations is a further use case – in 2009, a study of selected research papers published online over a ten-year period found that almost a third of the citations no longer linked to their original resources (Hanief Bhat, 2009). Citing versions from a web archive increases the longevity of the citation and its value for future readers. Initiatives such as the International DOI Foundation and DataCite address this issue and can provide part of the solution to it. Future trends in the development of web technologies and in academic research communities are likely to lead to more complex use cases as researchers look beyond the intellectual content of individual web pages and begin to treat web archives as massive temporal datasets for query and analysis.

These drivers and possibilities are just the tip of the iceberg (Leetaru, 2012; Brown, 2006). Perhaps the question then is not so much ‘why archive websites?’ as ‘why not?’.

## 1.2. The challenge of web archiving

Capturing and archiving an individual, small and simple website can be relatively straightforward; it can be copied or downloaded from the server and stored offline with relatively little technical expertise. Web archiving at scale and for the long term is a more complicated affair. Capturing large and complex sites on a recurring basis, whilst maintaining and clearly identifying the relationships between different versions of a site and simultaneously managing the artificial boundaries that inevitably occur in an ‘extracted’ collection, requires a more complicated solution. Add in the sheer scale of a ‘national domain’ collection and the long-term problem of technological obsolescence, and the challenge increases again.

Technical solutions have been developed to address many of these challenges, making it easier for smaller organizations to successfully embark on a web archiving initiative. But some issues have yet to be fully resolved. Temporal coherence is one such example: a large website may have changed before a web archive has even finished capturing a copy of it. The temporal integrity of the site is therefore unclear. Similarly, links within different sites in a web archive may lead to sites hosted at different times, skewing the user's perception of what information was available when. The rapid pace of change means that web archives are forever shooting at a moving target. Web archiving solutions need to work fast, and to capture as much material as possible in order to provide future users with an 'authentic' experience.

The form of an 'authentic' experience however, is anything but clear, and some academic issues remain. What for example, is an 'authentic' archived website? Is an archived website still authentic if some of the links are broken, or content missing? How can criteria for authenticity even be determined when sites do not exist as static objects but are generated dynamically and rendered differently for different users? These issues are still being explored.

Other challenges and issues range from legal uncertainties (such as copyright, data protection and libel) to quality assurance (ensuring that all necessary files have been captured and will render), long-term digital preservation (that they will continue to persist and render far into the future, despite changes in technology), and simply ensuring that web archiving technology keeps pace with the ongoing evolution of the web. These are discussed in more detail in Section 4.

### 1.3. Key web archiving initiatives

Most large-scale web archiving efforts to date have been driven by national libraries and archives, with one notable exception: the Internet Archive. The Internet Archive is home to the largest known collection of legacy web content, with over 2 petabytes of compressed data and over 150 billion captures of websites.<sup>1</sup> It was founded in 1996 as a non-profit organization with the mission of providing 'universal access to all knowledge', at around the same time as the National Libraries of Australia and Sweden began their web archive collections (PANDORA and Kulturarw3 respectively). In 2000, the National Library of Sweden joined forces with the four other Nordic national libraries to form the Nordic Web Archive (Brygfjeld, 2002). A number of other national heritage institutions followed suit and in 2003, recognizing that a global web archiving solution required a global collaborative effort, 11 of them joined forces with the Internet Archive to form the International Internet Preservation Consortium (IIPC, 2012a).

Founding members of the IIPC include the British Library, the Bibliothèque nationale de France, the Library of Congress, the National Library of Australia, and the National Library of Sweden. The Consortium has since trebled in size, and members have played key roles in defining a standard architecture and development of new web archiving tools whilst simultaneously developing their own collections. Another major contribution is the development of standards, exemplified by the

---

<sup>1</sup><http://archive.org/projects/>

WARC standard, alongside R&D work around metadata usage, workflows, access, and quality assurance.

Smaller, practical collaborative efforts such as the UK Web Archiving Consortium (UKWAC) and the aforementioned Nordic Web Archive (Bailey & Thompson, 2006; Brygfjeld, 2002) also have value in addressing and delivering solutions at a more local level. Though UKWAC was disbanded in 2008, collaborative web archiving efforts in the UK remain on the agenda, and since 2009 have been coordinated by the Digital Preservation Coalition (DPC) through the DPC Web Archiving & Preservation Task Force (DPC, 2012). Membership of the Taskforce is open to all associate and full members of the DPC, regardless of their size and the scale of their web archiving interests.

Recent years have seen commercial web archiving services appear. Two key services offered by major international players are the Archive-It service offered by the Internet Archive, and the Archivethe.net web archiving service offered by the Internet Memory Foundation (IMF). Other commercial web archiving services include those provided by, for example, Hanzo Archives and Reed Archives, both of which cater to the recordkeeping and archiving requirements of a wide range of clients.

A significant level of funding for web archiving R&D has been made available by the European Commission in recent years. Several major projects have been funded to either focus on, or include, web archiving:

- The SCAPE project (2011 – 2014), developing scalable long-term preservation services for complex heterogeneous and large-scale digital content including web archives. SCAPE runs a Web Content Testbed for testing preservation tools and issues specific to web archives.
- LAWA (Longitudinal Analytics of Web Archives, 2010– 2013), developing tools specifically to analyse heterogeneous Internet data at a massive scale.
- ARCOMEM (2011–2014), focusing on archiving and preservation of the social web, particularly social networks.
- BlogForever (2011–2013), which aims to develop archiving and preservation functionality specifically for content from online blogs.
- LIWA (Living Web Archives, 2008–2011), which explored ‘next generation’ tools for archiving web content.

## 1.4. Technical approaches to web archiving

Technical approaches to web archiving vary according to the scale of the operation. For large-scale archiving, Masanés (Masanés *et al.*, 2006) identifies three main technical approaches:

1. Client-side archiving
2. Transactional archiving
3. Server-side archiving

Client-side archiving is the most widely used approach. It is scalable, requires little input from the content owner, and can be highly cost-effective. Web crawlers such as Heritrix or HTTrack act as clients (i.e. browsers) and use the HTTP protocol to gather content responses delivered directly from the server. Web crawling technology was originally developed for indexing purposes, but has been adapted for web archiving to collect content in a modified fashion (Masanés *et al.*, 2006). The key adaptations are improved resource discovery and better regard for preservation of original content. The crawler follows a ‘seed’ instruction (typically a URL) and crawls all links associated with that seed to a specified depth, capturing copies of all available files. These are returned to the operator in a pre-determined form and processed accordingly.

Transactional archiving addresses a different use case, specifically the capture of client-side transactions rather than directly hosted content. It supports the growth of more comprehensive collections that record user access to site content, based on actual client/server transactions over time. Storing and archiving content associated with all unique HTTP request/response pairs, transactional archiving records the content that was presented to a user on a given date and time. Transactional archiving requires implementation of code on the web server hosting the content, so is mainly used by content owners or hosts rather than external collecting organizations.

In a similar fashion, direct server-side archiving also requires active participation from publishing organizations. Files are copied directly from the server without recourse to the HTTP protocol, but issues are often encountered in generating a working version of the content, particularly when absolute links have been employed, when content is database driven, or when creating a similar hosting environment to that of the original live website. Nonetheless, it can be a useful way to gather content otherwise missed by crawlers.

A fourth approach has been explored in recent years that uses RSS feeds to identify and pull externally hosted content into a web archive, rather than deploying a crawler. RSS feeds provide a trigger or alert that new content has been published, and can be used to ensure content is not missed in the intervals between periodic snapshot crawls. The ArchivePress and BlogForever projects both explored technical options for gathering blog content using RSS feeds (Pennock & Davis, 2009; Rynning *et al.*, 2011), and the WebDam project proposed their use in conjunction with all manner of websites (Oita & Senellart, 2010). Whilst BlogForever and WebDam both utilize crawlers to gather content, the ArchivePress project harvested content solely from RSS feeds and excluded the more general files associated with the look and feel of the site. Of the three, only ArchivePress tailored the approach based on the perceived important characteristics of the website.

Each of these four approaches captures slightly different content and requires different levels of technical skills and/or access to original content. Web archiving programmes that are primarily concerned with collecting external content often select client-side technology as it offers the broadest support. Server-side is most frequently used where content cannot be collected over HTTP (for example, a database or maps servers). Transactional archiving is used in the specific case where it is the user’s actions and entered data that are required, for example when archiving financial/commercial transactions for non-repudiation. Scale can also be an issue. ArchivePress, for example, is particularly well suited to establishing small collections focused on a specific theme or

topic, whereas large-scale archiving is better served by the flexibility, efficiency, and comprehensiveness offered by client- or server-side archiving. Smaller institutions wishing to establish collections using the latter approaches may find benefit in utilizing a commercial or third-party service if the technical capacity in their own organization is lacking. Collecting organizations should select that which is best able to meet their requirements with their available resources in the most efficient and effective manner.

## 2. Issues

### 2.1. Legal

Legality is often the biggest non-technical issue faced by web archives. Do they have the legal right to take copies of content and provide access independently of the original site and without explicit permission of the owner, or is that a breach of the owner's copyright? Some websites display clear licences or copyright information, such as Creative Commons or Crown Copyright, that go some way to answering that question. However, in most cases the answer largely depends on a) the country concerned and b) the remit of the collecting institution.

In the UK, legal deposit legislation for electronic publications and websites has not yet been implemented. This means that no single collecting institution in the UK can yet harvest the entire UK domain without risking copyright infringement. When the legal deposit legislation is passed (expected later in 2013), UK legal deposit libraries will be granted the right to gather and provide access to copies of all websites published in the UK domain. Until such time, the UK Web Archive (coordinated by the British Library) operates a selective, permissions-based web archiving model. The UK Government Archive maintained by The National Archives (UK) is slightly different – with a smaller scope and clearer statutory powers (namely a clear legal mandate from the Public Records Act and a significant proportion of Crown Copyright material in its collection), the Archive does not need to request permission for most of the material in its collection (The National Archives (UK), 2012). Despite this, issues can still arise and both institutions have a clear notice and takedown policy.

In the USA, legal experts have theorized that cases involving Google crawls could work as precedents for web archiving, arguing that 'the library website archives services a preservation and scholarly function that provides a significant benefit to the public' – it is done for a different purpose, which renders the use transformative and therefore 'fair use', though this has yet to be put to the test (Band, 2012). The Internet Archive has no explicit legislative permission to archive websites but operates on a 'silence is consent' approach, crawling the web anyway and taking down websites should the owner request it. The Library of Congress on the other hand, archives on a permissions basis (Grotke, 2011). Elsewhere, the situation varies from country to country: some have passed legal deposit legislation but restrict access solely to the reading rooms. In others there is no legal deposit legislation and collections are either built solely on a selective and permissions basis or are held in a 'dark archive' inaccessible to the public. A survey carried out by the British Library in 2011 reported that by June 2012, 58% of national libraries expected legislation would be in place to support domain-wide web harvesting (Brindley, 2011).

Copyright can pose further problems when additional or altered copies of the work are generated as part of a long-term preservation strategy (Brown, 2006), as some copyright legislation does not explicitly permit copying of works for this purpose. The legality of such action should be clarified either through legal advice or through appropriate discussions and licensing with the content owner. Libel, accusations of plagiarism, and data protection or privacy must also be considered, and collection approaches to each clearly stated in an organization's web archiving policy, based on appropriate legal advice.

## 2.2. Selection

Selection policies for web archives are typically consistent with broader, organizational collection policies. There are, in general, two main types of collections, both of which are clearly scoped:

- Domain collections – these often attempt to gather all websites associated with a particular country (i.e. national domain collections). This may include not only websites ending with the national domain suffix (e.g. .fr or .pt), but also websites hosted in that country with a different domain suffix, or websites hosted abroad whose content focuses on the collecting nation.
- Selective collections – individual websites are selected for inclusion in a collection based on their relevance to the collecting body's collection policy. These often take the form of 'special collections' which group together websites on a given theme or subject. Event-based archiving is a type of selective archiving that generates special collections in response to a specific event, such as national elections or the Olympics.

The main issue in establishing scoped collections is the artificial limits they impose, even at a national domain level. The Internet does not respect collection and national boundaries! Sites in these collections will frequently link to other sites that are not captured as part of a collection, and this can be frustrating for users who inevitably then encounter broken links.

These different types of collections have their own strengths and weaknesses:

- Domain collections are potentially the most comprehensive, yet current limitations in web archiving technology mean that websites in domain collections are often incomplete: files may not been harvested or have been harvested but are not rendering properly, or the full extent of the site has not been captured. The bigger and more complex the site, the more likely it is to be incomplete. However, the sheer scope of domain collections means that relationships with other sites and external linked content are more likely to be maintained than in a site archived as part of a selective collection.
- Selective archiving focuses resources on sites considered to be particularly valuable and enables capture within a certain collecting remit. This value measure, whilst contentious, generally requires that the quality of archived sites reaches a minimum level. Sites are therefore more likely to be 'complete', i.e. with all files present and rendering correctly, even though links to external sites are more likely to be broken. Such quality assurance requires additional resource and is discussed in more detail in Section 4.4.

A further potential weakness of selective archives is their possible or unintentional and unacknowledged selector bias. Selection of sites is commonly a manual process that reflects the particular interests or knowledge of the person(s) choosing sites for the collection. The sheer size of the Internet, the number of websites hosted and the speed at which information can be published, all make it very difficult for manual selectors to keep abreast of new sources, especially for event-based collections. As a result, selective collections run the risk of being unintentionally

biased and their research value constrained. Retaining information about the selectors and their interests can help to alleviate the problems caused by such bias.

Overlapping selections may be an issue for some collecting institutions, particularly if of the same nationality. What is the cost-benefit of two (or more) institutions archiving the same site? How does that benefit the user? What impact does it have on the website owner? A clear archiving policy can go some way to controlling these issues and addressing them in a mutually beneficial manner.

### 2.3. Limitations of current crawlers

Most web archives deploy web crawlers to crawl and harvest copies of web content. Crawler technology has come a long way over the past decade, yet some limitations remain in the types of content crawlers are able to easily capture. Problematic content includes:

- Database/dynamically driven content (i.e. web pages that are generated via a database in response to a request from the user);
- Streamed multimedia files;
- Content accessible only via local site searches – script code is almost impossible for crawlers to analyse;
- Password-protected content – crawlers can deal with this if supplied with the password, but without the password the content is difficult to access;
- Some types of Javascript-driven menus – e.g. when URLs are generated by dynamic mechanisms.

Some of this content is often referred to as the ‘deep web’. The ‘deep web’ contains content which is difficult for crawlers to see and therefore access. Dynamically generated and password-protected content both fall into this category.

Other issues that can halt or prevent a crawler from making progress include operational limits on crawl size (where a crawl scope exceeds the amount of crawler memory available to store discovered hosts or scheduled URLs) and crawler traps (e.g. shopping carts or online calendars with dynamic pages and no fixed end date). Note, however, that an operational limit is not a limitation of the crawler but a practical measure to limit crawl size. The IIPC ‘Future of the Web’ workshop paper provides more detail on these and other technical issues associated with crawlers (IIPC, 2012a).

Research and development in this area is ongoing to address these limitations in future generations of crawlers.



## 2.4. Authenticity, Integrity and Quality Assurance.

What should an authentic archived version of a website look like? Should it be identical in all essential respects to the original 'live site'? Capturing and rendering an identical copy of a site may have been an admirable goal in the early days of the web, when websites were simpler and sites displayed messages informing users which browser their site was optimized for (IIPC, 2012a), but as we move towards personalized browsing experiences it is becoming increasingly difficult to identify what comprises an 'original site' and how it should look in the archive. Not only do different browsers affect the overall look and feel of a website, but even the content presented to the visitor changes. The concept of an 'original site' is becoming somewhat meaningless. How then can an archived version of a site be validated as 'authentic'? The 'significant properties' concept is one useful way for validating the success of a preservation approach, identifying key aspects of original sites relating to content, context, appearance, structure, and behaviour (e.g. Knight & Pennock, 2008). Ball (2010) provides a useful summary of important textual, interactive, dynamic and appearance characteristics of websites. Consideration must also be given to what the institution has set out to capture – the 'preservation intent' – and whether that intention is clear to the user (Dappert & Farquhar, 2009). Assessment of this sort is not simple, but it enables an organization to gauge the authenticity requirements against which a captured site can be validated.

Validation commonly takes place within a Quality Assurance (QA) process. Manual QA may be carried out by a trained specialist to assess a) what the crawler has collected and b) how the archived website renders in a standard browser. This ensures that not only have the target files been captured, but also that they render acceptably and in accordance with the preservation intent established by the collecting institution. Manual, visual QA is time consuming. Automated QA tools offer a more efficient way to approve crawls within a very large-scale environment, though they cannot drill down to the same level as a trained human eye. Tools currently in development look instead at a number of key indicators that may indicate problems with the crawl, such as long URIs, obvious crawl errors (noted in the crawl logs), missing links, data download size and unknown MIME types (Hockx-Yu and Woods, 2012; Raditsch, 2012).

## 2.5. Archiving Web 2.0

Web 2.0 sites are commonly rich in JavaScript applications which, as noted above, can cause problems for crawlers. Just as pertinent is the question of whether Web 2.0 sites are sufficiently dissimilar to 'traditional' websites that they should be archived in a slightly different way. For example, should the 'History' pages of a wiki make a difference to crawl frequency, given that they enable the crawler to capture not just the live but also historical versions of the site? Perhaps capture is not so time-dependent as with more traditionally structured sites (Pinsent, 2009) and the standard crawl frequency can be lessened. Blogs are similar - each new post is an addition to the site and older posts usually remain available in the blog archive rather than being overwritten. The JISC PoWR Handbook further points out that the 'fluidity' which often characterizes Web 2.0 content can make it difficult to identify the point at which content has been completed and is therefore ready for archiving (ULCC & UKOLN, 2008).

Social networking sites pose a different challenge. Twitter, for example, is not just about tweets, but about the conversation. To archive a single Twitter page therefore is to archive only one side of the conversation. How do you establish the boundaries for a coherent Twitter collection – is it necessary to capture all @replies to a Twitter account, as well as the account itself? Should user profiles for the @replies also be archived to provide a (small) amount of contextual information? Given the importance of links on Twitter, should all links tweeted from a target account also be archived? How can an archive ensure temporal consistency between tweeted links and the content of the linked site, especially given the very short half-life of a link on Twitter? Setting the boundaries of a social networking site is less straightforward than it may at first appear.

A further complication is the issue of permissions. Web 2.0 sites such as wikis, blogs (with comments), social networking sites and media-sharing sites typically contain a significant amount of multiple user generated content. For permissions-based collections, any site with user generated content poses the challenge of either asking the site owner to provide clearance, or collecting permission from all contributors. This is a time-consuming and sometimes near impossible task.

Although many of these problems have yet to be solved, it is still arguably worth attempting to capture some of this content before it is lost.

## 2.6. Temporal coherence

Web archiving has a fascinating temporal dimension. The greater the time period spanned by the archive, the greater its temporal value. Unfortunately, it is also the case that the bigger the archive and the sites in it, the bigger the risk of fracture to its temporal coherence. Temporal coherence is described as ‘a property of a set of archival Web pages, indicating that there was a point in time at which all the archived pages were live simultaneously on the Web’ (Ball, 2010). Temporal incoherence occurs when in the time taken for a crawler to finish crawling a website, parts of the site have been updated and content from the top levels of the seed URL (e.g. the homepage) no longer temporally matches those from the lower levels. This is a potential issue even for websites of moderate size (say, 2+MB). For a domain level collection, it becomes even more of a challenge, as depending on the size of the domain, a full crawl may take days or weeks to complete. The resulting collection cannot be considered as a representative copy of the web (or websites) on a given day, but only over a given period. This matters to future researchers who wish to know what information was available to a historical user at a given point in time. Tools to ensure temporal coherence at a single site level are emerging (Mazeika *et al.*, 2010) though for many users and collecting institutions it remains a conceptual and practical challenge.

## 2.7. Viruses and malware

The term ‘malware’ is commonly used as a catch-all phrase for unwanted software designed to infiltrate a computer, operating system, or application such as a web browser or a word processor, without the owner’s informed consent. It includes but is not limited to viruses, Trojan horses, worms, rootkits, dishonest adware, spyware, key-loggers, crimeware, cookie trackers and other

malware. The impact of each type is different though Anti-Virus (AV) software typically identifies all of these sources as infections.

Whilst highly undesirable for most contemporary web users, malware is a pervasive feature of the Internet. Many archives choose to scan harvests and identify malware but prefer not to exclude or delete them from ingest into their repositories, as exclusion threatens the integrity of a site and their prevalence across the web is a valid research interest for future users. Furthermore, AV software typically errs on the side of caution and is known to result in false positives, so exclusion of material on the basis of a positive AV software scan could result in unnecessary omissions. The retention of malware may however conflict with organizational policy, particularly in large-scale organizational repositories. Whatever the approach, hosting institutions should ensure users are aware of it and can take steps if necessary to protect their PCs.

## 2.8. De-duplication

The term 'de-duplication' refers to the elimination of multiple copies of identical content so that fewer copies are kept. It comprises both de-duplication of technically identical content (i.e. the bits and bytes) and the more nuanced view of de-duplication at an intellectual content level. De-duplication is an important issue for web archives as multiple copies of identical content are frequently collected in different captures of a website over time. Two aspects give it particular significance:

- The volume of content duplicated, suggested to be on average around 25% (Gomes *et al.*, 2006) though this clearly depends on the crawl frequency;
- The scale of web archives and the sheer amount of storage space required to support duplicated content.

In some cases de-duplication may be implemented until only one copy of a file remains. In other cases there may be some benefit in less de-duplication, leaving more than one copy of a file in the archive but still reducing the number of copies kept overall. This reduces storage costs but retains a minimum level of duplication as good practice in case of subsequent problems or file corruption. De-duplication is supported in the WARC archival storage format (see Section 5).

There are cases where de-duplication is not desirable as it conflicts with the preservation intent and business case of the collecting institution. For example, web archiving for capture of legal records should avoid de-duplication as each instance of a site needs to be able to stand alone and each object shown captured alongside the rest of the objects in the site. If this is not done, the opposing counsel may reject the capture on the grounds of spoliation.

## 2.9. Search limitations

Searching a web archive is different from searching the live web (Niu, 2012; Stack, 2005). In the absence of a de-duplication strategy, search results from a web archive are often distorted by the presence of multiple copies of identical content harvested in different crawls. Addressing this distortion poses a significant challenge.

Ranking is also different from the live web, particularly in selective archives. Up until recently, very little was known about optimizing ranking in search results for web archives (da Costa, 2011), which made it difficult for new users to understand search results and find relevant content. New search-indexing technology for web archives goes some way in addressing this, and operational indexing of collections utilizes very large-scale technology such as SOLR (see Section 6) and/or HIVE, commonly used in conjunction with Hadoop, an open-source software framework that supports efficient large-scale data processing. In the meantime, institutions can provide explicit guidance for users on how to search their collections (The National Archives (UK), 2011).

## 2.10. Long-term preservation

Technical approaches for long-term digital preservation address the challenge of technological obsolescence by ensuring content can be reliably accessed and rendered in the future despite changes to the original hardware/software environment. Rendering requirements for some types of files are relatively loosely specified – straightforward HTML (Hyper Text Markup Language) files, for example, can usually be rendered well by most browsers running on a range of hardware/software platforms. Other types of files are more dependent upon a specific environment. In general, the older the file, the more likely there are to be problems rendering it accurately on a modern computing platform. Thompson (2008) gives an excellent summary of the two main technical problems for long-term preservation of web archives:

1. The sheer complexity of the vast range of formats published on the web. These must not only all be captured, but also a mechanism developed for maintaining access to them;
2. The complex relationships between files that comprise webpages and websites. The structural relationships and active links between different files and components of a webpage and website must not only be captured and made to work in an independent archival environment, but those relationships must also be maintained over time. This is increasingly difficult if a migration strategy is used and the filenames changed. At a domain level, this is even more of a challenge.

A migration on request strategy, whereby a specially developed web browser automatically migrates files in legacy formats when requested by the user, could potentially address both issues.

A study recently published by the UK Web Archive, based on a format assessment of a 15 year UK domain collection harvested by the Internet Archive, suggests that the range of formats is perhaps

not as great an issue as has previously been thought, and that web formats for text and images may not become obsolete as fast as had been previously feared (Jackson, 2012). However, the study also clearly illustrates the extent to which the web community embraces new versions of formats and markup languages. The ability of contemporary web software to reliably render old versions of still current formats has yet to be fully assessed.

Other tools such as PLATO and DROID also support preservation of web archives and have already been used at small scale (Kulovits, 2009) with large-scale tests pending (Raditsch, 2012). Recently published data suggests however that it is overly time consuming and perhaps unnecessary to run validation and identification tools on domain level collections (Jackson, 2012). The WARC standard has much improved support for long-term preservation compared to its predecessor, ARC, and a new tool, JhoNAS<sup>2</sup>, is being developed to foster WARC usage in scalable web archiving workflows using Jhove2 and NetarchiveSuite (Clarke, 2012). The IIPC Preservation Working Group is actively pursuing a range of preservation-related projects (IIPC 2012b) including the WARC Tools Project<sup>3</sup>. The KEEping Emulation Portable (KEEP) project has developed an emulation framework and services with the National Library of the Netherlands (KB) that may enable preservation and accurate rendering of archived websites over time (KEEP, 2013). Overall, good progress is being made on individual aspects of the long-term challenge, though a unified solution for preservation across the entire lifecycle has yet to be definitively implemented.

For more information about technical long-term digital preservation strategies for web archives, readers are referred to more extensive discussions and strategies featured elsewhere (e.g. Brown, 2006; Day, 2003).

---

<sup>2</sup> <http://netpreserve.org/projects/jhonas>

<sup>3</sup> <http://netpreserve.org/projects/warc-tools-project>

### 3. Standards

This section explores web standards; standards and guidelines for web archiving; and standards for long-term preservation.

HTML and related HTML/XML technologies are the core standards upon which the interoperability, and thereby success, of the Internet is based. HTML was the initial language of the Internet. It has led to the development of several related technologies, including XML (eXtensible Markup Language) and the XML family including XSLT (eXtensible Language Stylesheet Transformations), which is a language for transforming XML documents into other documents and/or formats. All of these are common on the web. HTML and XML are well-defined de jure web standards, maintained and excellently documented by the W3C. The most recent version of HTML is HTML5, which has potential to make the harvest of multimedia material easier (IIPC, 2012a). Other standards that web crawlers are most likely to encounter include CSS (Cascading Style Sheets), Javascript, and HTTP. HTTP is the protocol for exchanging hypertext and is essential for communicating data across the web.

A different kind of web standard is the robots.txt protocol. This is commonly used by webmasters to indicate whether they permit crawling by automated bots. It can be useful in keeping web crawlers out of crawler traps, but it is a default setting with some website creation software and may be set without the explicit knowledge of the website owner. The robots.txt protocol is advisory and crawlers can be set to ignore it should adherence to the protocol prevent the crawler from capturing material required by the web archive.

ISO 28500:2009 and ISO Technical Report 14873 support core web archiving activities. ISO 28500:2009 is more commonly known as the WARC standard. WARC was designed by IIPC members as a standard method for packaging together multiple files crawled from one or more websites, maintaining and describing the relationships between them whilst allowing the addition of key crawl and archiving metadata. It is based upon the ARC container format for web archive content previously developed by the Internet Archive and was designed to better support long-term preservation, as data and structure to support long-term preservation in ARC was minimal. WARC is extensible and files can accommodate related secondary content, such as assigned metadata, abbreviated duplicate detection events, and later-date transformations. WARC was approved as an ISO standard in May 2009. The ISO Technical Report 14873 on statistics and quality issues for web archiving (due to be published in 2013) identifies a number of meaningful collection development statistics and quality indicators relating to ongoing accessibility of content.

In the UK, the Central Office of Information (COI) Standard on Archiving Websites, written by The National Archives (UK), provides best practice guidance to webmasters on website creation and maintenance as well as process guidelines to support The National Archives (UK) in archiving government websites (COI, 2008). It requires that websites of all central government agencies are able to be archived by The National Archives (UK) three times a year and before sites are terminated. A similar paper on managing URLs requires good practice from webmasters on persistent and meaningful URLs (COI, 2009). Though the COI has now closed, the guidance remains valid and the standards are now managed by the Government Digital Service (GDS). Standards for

persistent identifiers have a clear role to play in web archives, supporting the use of redirection technology (Spencer *et al.*, 2008) and ensuring that content can consistently be located despite possible changes to the underlying web archiving infrastructure. Niu (2012) notes at least four web archives that provide persistent identifiers for archived web pages, using customized URIs. WebCite on the other hand, is an online web referencing service that uses the Digital Object Identifier (DOI) standard when possible and under certain circumstances can assign DOIs to its archived copies (WebCite, 2012).

An ISO standard with particular support for long-term preservation is ISO 14721:2012, otherwise known as the Open Archive Information System (OAIS) Reference Model. The OAIS model describes a general framework for the long-term preservation of digital content within a repository. Though not specific to web archives, it provides a useful idealized workflow for storage and management of complex digital content at scale without relying upon a specific technical solution or tool.

No section on standards would be complete without at least a brief mention of metadata. Different metadata standards can be utilized in a web archiving metadata profile, depending on the needs and requirements of the collecting institution (Brown, 2006). Marc 21, ISAD(G) and Dublin Core can all be (and have been) used to record descriptive metadata about web archives, with resource discovery further enhanced by use of Library of Congress Subject Headings (LCSH) and/or the Dewey Decimal Classification (DDC). Similarly, both METS and PREMIS can and have been used within WARC containers to record additional information about web archives: PREMIS for preservation metadata, and METS as a wrapper for descriptive, administrative and structural metadata (Enders, 2010).



## 4. Software

This section introduces some of the technical solutions currently available or in development for web archiving. It identifies a number of key software options that readers should be aware of, particularly for large scale collections. A number of other smaller-scale or application-specific services are available, such as Diigo, Zotero and Mendeley. These are useful for organizations or individuals that wish to establish small collections for personal and/or scholarly use.

### 4.1. Integrated systems

A small number of integrated systems are available for those with sufficient technical staff to install, maintain and administer a system in-house. These typically offer integrated web archiving functionality across most of the life cycle, from selection and permissions management to crawling, quality assurance, and access. Three are featured here.

#### 4.1.1. PANDAS

PANDAS (PANDORA Digital Archiving System) was one of the first available integrated web archiving systems. First implemented by the National Library of Australia (NLA) in 2001, PANDAS is a web application written in Java and Perl that provides a user-friendly interface to manage the web archiving workflow. It supports selection, permissions, scheduling, harvests, quality assurance, archiving, and access.

PANDAS is not open source software, though it has been used by other institutions (most notably the UK Web Archiving Consortium from 2004 to 2008). The current version is V3, released in 2007 and representing a significant re-write of the V2 code to improve performance and enhance sustainability. Component modules include:

- *PANDAS* – as the core component module, PANDAS encapsulates the functionality required to provide a web-based user interface for managing the workflow activities.
- *Crawler* – HTTrack is used, directly connected to PANDAS.
- *PANDORA* – this component creates the public interface through title and subject listings, as well as Title Entry Pages (also known as TEP pages) for archived resources. These TEP pages display resources in a predefined format and are generated on the fly.

Other modules manage related functions, such as access restrictions and reports. One of the main issues with PANDAS is its use of HTTrack, which is not optimized for web archiving and is discussed further in Section 6.3. Other community concerns relate to ongoing support and active development. Concerns over the PANDAS use of WebObjects, a Java web application development framework from Apple Inc. for which support was withdrawn in 2007, have been somewhat alleviated recently by the emergence of the WebObjects Community Developer Association.



PANDAS is used by the NLA for selective web archiving, whilst the Internet Archive supports their annual snapshots of the Australian domain.

#### 4.1.2. Web Curator Tool (WCT)

The Web Curator Tool is an open source workflow tool for managing the selective web archiving process, developed collaboratively by the National Library of New Zealand and the British Library with Oakleigh Consulting. It supports selection, permissions, description, harvests, and quality assurance, with a separate access interface. WCT is written in Java within a flexible architecture and is publicly available for download from SourceForge under an Apache public licence. It is comprised of the following main independent modules:

- WCT Core Component – this includes the web server, scheduler and other central components, for example authorization and identification, scheduling, a user (i.e. archivist) interface, and a harvest agent.
- Harvester – Heritrix is integrated into Core as a harvest agent. It is also used as an independent application to crawl and harvest sites on instruction from the archivist.
- Wayback – Wayback is an independent application from the Internet Archive that provides front end user access to approved archived content.

The most current WCT release is V1.6 (December 2012). WCT issues are managed by the main developers, though raised issues often relate to crawler limitations rather than WCT per se. The WCT website is the hub for the developer community and there are active mailing lists for both users and developers. The highly modular nature of the system minimizes system dependencies.

#### 4.1.3. NetarchiveSuite

NetarchiveSuite is a web archiving application written in Java for managing selective and broad domain web archiving, originally developed in 2004 by the two legal deposit libraries in Denmark (Det Kongelige Bibliotek and Statsbiblioteket). It became open source in 2007 and has received additional development input from the Bibliothèque nationale de France and the Österreichische Nationalbibliothek since 2008. It supports selection, permissions, scheduling, harvesting, quality assurance and access, and is freely available under the GNU Lesser General Public License (LGPL). The suite is split into four main modules:

- Core Netarchive suite – this consists of several modules. A common module provides the overall framework and utilities for the suite, plus an interface between the different modules (which include harvest and two access modules (user/archivist) as well as monitoring and archiving modules).
- Harvester module – Heritrix, wrapped in a customized interface for scheduling, configuring and distributing crawls.
- Access modules:

- one for archival management of the crawl
- a second for access (Wayback).

The highly modular nature of the system enables flexible implementation solutions (Sørensen *et al.*, 2012).

## 4.2. Third party/commercial services

Third party commercial web archiving services are increasingly used by organizations that prefer not to establish and maintain their own web archiving technical infrastructure. The reasons behind this can vary widely. Often it is not simply about the scale of the operation or the perceived complexity, but the business need and focus. Many organizations do not wish to invest in any skills or capital that is not core to their business. Others may use such a service to avoid capital investment. Moreover, organizations are increasingly moving their computing and IT operations into the cloud, or using a SAAS (Software as a Service) provider. Web archiving is no exception. From a legal and compliance perspective, third party services are sometimes preferred as they can provide not just the technology but also the skills and support required to meet business needs. This section introduces some of the third party services currently available but is of course a non-exhaustive list, and inclusion here should not be taken as recommendation.

Archive-It is a subscription web archiving service provided by the Internet Archive. Customers use the service to establish specific collections, for example about the London 2012 Olympics, government websites, human rights, and course reading lists. A dedicated user interface is provided for customers to select and manage seeds, set the scope of a crawl and crawl frequency, monitor crawl progress and perform quality assurance, add metadata and create landing pages for their collections. Collections are made public by default via the Archive-It website, with private collections requiring special arrangement. The access interface supports both URL and full text searching. Over 200 partners use the service, mostly from the academic or cultural heritage sectors. The cost of the service depends on the requirements of the collecting institution.

Archivethe.Net is a web-based web archiving service provided by the Internet Memory Foundation (IMF). It enables customers to manage the entire workflow via a web interface to three main modules: Administration (managing users), Collection (seed and crawl management), and Report (reports and metrics at different levels). The platform is available in both English and French. Alongside full text searching and collection of multimedia content, it also supports an automated redirection service for live sites. Automated QA tools are being developed though IMF can also provide manual quality assurance services, as well as direct collection management for institutions not wishing to use the online tool. Costs are dependent upon the requirements of the collecting institution. Collections can be made private or remain openly accessible, in which case they may be branded as required by the collecting institutions and appear in the IMF collection. The hosting fee in such cases is absorbed by IMF.

The University of California's Curation Centre, as part of the California Digital Library, provides a fully hosted Web Archiving Service for selective web archive collections. University of California departments and organizations are charged only for storage. Fees are levied for other groups and consortia, comprising an annual service fee plus storage costs. Collections may be made publicly available or kept private. Around 20 partner organizations have made collections available to date. Full text search is provided and presentation of the collections can be branded as required by collecting institutions.

Other private companies offer web archiving services particularly tailored to business needs. Whilst it is not the purpose of this report to provide an exhaustive list and analysis of all commercial offerings, some examples may be useful. Hanzo Archives, for example, provide a commercial website archiving service to meet commercial business needs around regulatory compliance, e-discovery and records management. The cornerstone of their product suite is the Hanzo Enterprise solution, which uses their own commercial web crawler. Hanzo Archives emphasize their ability to collect rich media sites and content that may be difficult for a standard crawler to pick up, including dynamic content from Sharepoint, and wikis from private internets, alongside public and private social media channels. (More details about the possibilities afforded by the Hanzo Archives service can be found in the Coca-Cola case study.) Similarly, Reed Archives provide a commercial web archiving service for organizational regulatory compliance, litigation protection, eDiscovery and records management. This includes an 'archive-on-demand' toolset for use when browsing the web (Reed Archives, 2012). In each case, the cost of the service is tailored to the precise requirements of the customer. Other companies and services are also available and readers are encouraged to search online for further options should such a service be of interest.

### 4.3. Crawlers

Web crawlers are used by the web archiving services listed above to crawl and harvest copies of websites. Web crawlers are not exclusive to web archives, and crawler software is also used extensively by a broad range of other applications and services, especially by search engines. This section focuses on crawlers used to archive websites.

HTTrack is an early and free open source offline browser and crawler written in C and C++. HTTrack writes files to disc in a way that mirrors the URL structure of the crawled site. However, HTTrack also re-writes links so that they point to the local HTTrack copy of the files, rather than to the live website. The original filenames are therefore lost. This has led many organizations to prefer the Heritrix crawler, now very widely used by web archiving institutions (Internet Archive, 2012). Originating from the Internet Archive, Heritrix is open source, written in Java, and can be configured to store harvests in either ARC or WARC containers. It supports both selective and domain crawls and can be adapted to collect streamed content including YouTube videos. Heritrix can also be configured to de-duplicate data harvested in previous crawls.

WGet is a free utility tool for downloading files from the web that can be configured to act as a crawler, downloading files from the web to the depth specified by the user. Awareness of WGet as a web archiving crawler has increased after usage by the ArchiveTeam. The most recent version of

WGet (v1.14) will deliver WARC output and store the request/response headers otherwise missed alongside 404s and redirects. This is a significant improvement on the previous version for web archiving purposes.

Third party services may use their own design of crawler, or adapt a publicly available crawler. The Miyamoto crawler designed by Hanzo Archives, for example, is a mature and highly capable web crawler that has been designed to capture web content considered problematic for other crawlers. The Memorybot crawler launched recently by Internet Memory was designed specifically for very large crawls and has shown an increase in the number of URLs successfully crawled when compared to other crawlers.

#### 4.4. Searching

Web archives are growing all the time. Whilst some collections remain relatively small, others are now very large – the archives d’Internet at the Bibliothèque nationale de France for example consists of around 300Tb of data, or 17 billion files. Large-scale archives require large-scale search capabilities. This section introduces some of the main tools used for searching and indexing web archives.

NutchWAX is a tool for indexing and searching web archive collections, used in conjunction with an access interface such as Wayback (see below). It is an adaptation of the open source web search engine Nutch, adapted to interrogate archives rather than the live web whilst acknowledging multiple captures of a single URL. It is written in Java, maintained in large part by the Internet Archive in conjunction with other users from the web archiving and Nutch communities. NutchWAX can be downloaded from SourceForge but development has more or less halted in recent years, mainly as a result of changes to the core Nutch search engine away from full text searching to crawling (for which the web archiving community has other tools). Concerns have also been raised about its less than optimum performance at very large scale, and poor support within Nutch for accented characters or multiple character sets (Binns, 2011).

SOLR is an alternative open source search platform, supported by the Apache Software Foundation. It is attracting increasing interest from web archiving institutions for its powerful search abilities, low cost, and fast performance. It is a highly scalable Java-based full text search platform that is often used with Hadoop, an open source software framework that supports efficient large scale data processing. Combining SOLR with Hadoop enables both efficient processing and delivery of search results. Both the Internet Archive and the UK Web Archive use this combination to great success. SOLR supports faceted searching, spell checking and field collapsing, all common functionality for live web searches and very popular with modern users. Out of the box functionality is good and it can be tailored to almost any type of application, though this should not be taken to mean implementation is straightforward or fast, nor is there formal support for upgrades or for less skilled developers.

## 4.5. Access

Access components and tools provide the front end for web archives and are the mechanism by which users explore and exploit the content. This section looks at some of the existing and emerging tools and interfaces for accessing web archives.

WayBack is an open source, Java implementation of the Internet Archive's Wayback Machine (Internet Archive, 2012). It is a common front end component, widely used by institutions to provide access to their web archive collections. The standard interface provides basic access to sites as captured by the crawler. Some implementations overlay a timeline on rendered pages so that users can clearly see which instance of a site they are exploring and navigate to other instances without having to return to the main interface.

The Memento framework allows users to access archived versions of websites they are already viewing on the live web, instead of via the access interface provided by web archives (Van de Sompel, 2009). It requires three main things:

1. implementation of Memento code on live sites;
2. implementation of Memento code on sites holding archival collections;
3. installation of the MementoFox plugin so that users can access the archived versions of sites from their browsers.

Memento has proven very popular in the digital preservation community and won the 2010 Digital Preservation Awards, awarded by the Institute for Conservation and the Digital Preservation Coalition, for its use of existing and widely deployed content negotiation tools to connect users with pre-existing archives (DPC 2010).

A number of analytical access tools are being developed with enhanced functionality to better explore and exploit the temporal nature of the collections. A recent paper from the Oxford Internet Institute, sponsored by the IIPC, suggested that visualization in particular is a key introductory access tool for large-scale web archives and that in an ideal future for web archives, 'much more powerful and effective tools for text search, information extraction and analysis, visualization, social annotation, longitudinal analysis, and sentiment analysis' will be needed (Meyer *et al.*, 2011). The UK Web Archive, for example, now provides an NGram viewer for users that returns search results within a graph to illustrate occurrences of the search terms in the web archive over time. Further tools have been prototyped, including tagclouds, subject hierarchy visualization, and postcode-based access (Crawford, 2011). An academic project in the USA has developed a further six visualization tools to work on the Archive-It collections, including a treemap, a time-tag cloud, and wordle (Padia, 2012).

The Web Continuity Service implemented by The National Archives (UK) for UK government websites offers an alternative yet eminently practical way for users of the live web to access

archived web material. When content is removed from government websites, redirection technology is deployed on the live sites to deliver users with that content straight from the web archive. Disappearing documents and broken links are no longer an inconvenience for users, and the web archive receives over 100 million hits a month as a result.

#### 4.6. Other options

Two software tools are being developed that do not fit easily into any of the categories above but illustrate particularly innovative solutions to specific web archiving challenges.

- SiteStory is a new solution from the Memento team that supports transactional web archiving (Van de Sompel, 2012). It is not a crawl-based solution; instead it requires code to be implemented directly on the web server. This allows it to capture the server's entire history and every version of a resource that is ever requested by a browser. The resulting archive can be accessed by Memento or via a Wayback interface.
- Twittervane is a new approach to selecting websites for theme- or event-based collections (Pennock, 2011). It addresses the potential issue of selector bias in a collection by producing a list of web resources relating to a given theme or event, ranked by the number of times they are shared on Twitter. Termed 'social selection', a prototype tool was developed by the British Library with funding from the IIPC and released as open source on Github (IIPC, 2012c). Further development is underway to develop a production version.

## 5. Case Studies

Three case studies have been selected to illustrate the different ways in which a web archiving solution may be implemented. The first is a national library with a self-hosted and self-managed, open source web archiving solution. The second is a web archiving service offered by a non-profit foundation and utilized by several small, medium or large institutions to host and manage their web archiving collections. The final study focuses on a commercial web archive hosted and managed by a third party commercial organization.

### 5.1. The UK Web Archive

#### Background

The UK Web Archive (UKWA) was established in 2004 by the UK Web Archiving Consortium in response to the challenge of a so-called 'digital black hole' in the nation's digital cultural Internet memory. The Archive contains UK and UK-related archived websites of scholarly and research interest, with a particular interest in sites that reflect the diversity of lives, interests and activities throughout the UK or which feature political, cultural, social and economic events of national interest. Archived sites are, wherever possible, made freely available through the UK Web Archive website.

#### Collaboration

The UK Web Archiving Consortium was originally a six-way partnership, led by the British Library in conjunction with the Wellcome Library, JISC, the National Library of Wales, the National Library of Scotland and The National Archives (UK). The Consortium was wound up in 2008 and only four of the original partners remain active contributors, all of which select and nominate websites using the features of the web archiving system hosted on the UK Web Archive infrastructure maintained by the British Library.

The British Library works closely with a number of other institutions and individuals to select and nominate websites of interest, including the Women's Library, the Live Art Development Agency, the Library and Information Services Council of Northern Ireland, the Quaker Library, the Mass Observation Archive, the Royal Archives, and MANX Heritage.

#### Approach

Legislation to enable legal archiving of websites from the UK domain without requiring explicit permission from website owners was drafted back in 2003, but the precise terms have been under debate for almost a decade. In the absence of this legislation, the Library operates on a permissions basis and focuses its attention on selective web archiving in line with the collection priorities outlined above. Partner organizations typically follow the same approach, with the notable

exception of the JISC who in 2011, in addition to archiving selected websites, also purchased a copy of the UK domain archive as collected by the Internet Archive from 1996 onwards. This domain level collection is stored by the Library on behalf of JISC but is not available for public browsing.

Selectively archived websites are revisited at regular intervals so that changes over time are captured.

## Technical solutions

The technical infrastructure underpinning the UK Web Archive is managed by the British Library. The Archive was originally established with the PANDAS software provided by the National Library of Australia, hosted by an external agency, but in 2008 the archive was moved in-house and migrated into the Web Curator Tool (WCT) system, which was collaboratively developed by the British Library, the National Library of New Zealand and Oakleigh Consulting. WCT is a workflow management tool that was designed specifically to support the selective web archiving process and to meet the Library's (and UKWAC's) needs for the UK Web Archive.

The migration necessitated a simultaneous change in web crawler, from HTTrack to Heritrix. The Heritrix crawlers are now configured to store files directly in WARC containers, though this has not always been the case and an ARC to WARC migration took place shortly after the PANDAS to WCT migration to ensure all archived sites are stored in as consistent a manner as possible. The Library has designed a METS-based SIP profile expressly for its web archives, which is used when ingesting web archives into the Library's long term store.

A customized version of the Wayback interface developed by the Internet Archive is used as the WCT front end and provides searchable public access to the collection.

## Access

The public interface at [www.webarchive.org.uk](http://www.webarchive.org.uk) provides access to all publicly available archived websites. Full text searching is enabled in addition to standard title and URL searches and a subject classification schema. The web archiving team at the library have recently released a number of visualization tools to aid researchers in understanding and finding content in the collection. For the selective archive, these include an NGram search, Tag clouds, and a 3D wall. Visualizations from the domain collection are also available, including an NGram search, link analysis, format analysis and GeoIndex.

## The Collection

Around 30 special collections have been established on a broad range of topics. Many are subject based, for example the mental health and the Free Church collections. Others document the online



response to a notable event in recent history, such as the UK General Elections, Queen Elizabeth II's Diamond Jubilee and the London 2012 Olympics.

Many more single sites, not associated with a given special collection, have been archived on the recommendation of subject specialists or members of the public. These are often no longer available on the live web, for example the website of UK Member of Parliament Robin Cook or Antony Gormley's One & Other public art project, acquired from Sky Arts.

## 5.2. The Internet Memory Foundation

### Background

The Internet Memory Foundation (IMF) was established in 2004 as a non-profit organization to support web archiving initiatives and develop support for web preservation in Europe. Originally known as the European Archive Foundation, it changed its name in 2010. IMF provides customers with an outsourced fully fledged web archiving solution to manage the web archiving workflow without them having to deal with operational workflow issues.

IMF collaborates closely with Internet Memory Research (IMR) to operate a part of its technical workflows for web archiving. IMR was established in 2011 as a spin off from the IMF.

### Collaboration

Internet Memory is, and has been, involved in several research projects to improve technologies of web-scale crawling, data extraction, text mining, and preservation. IMR works with several institutions in Europe, including the Institute for Sound and Vision in the Netherlands and the National Library of Ireland. Both IMF and IMR are involved in research projects that support the growth and use of web archives.

Partner institutions, with openly accessible collections for which the IM provides a web archiving service, include the UK National Archives and the UK Parliament.

### Approach

IMR provides a customizable web archiving service, Archivethe.Net (AtN). AtN is a shared web-archiving platform with a web-based interface that helps institutions to easily and quickly start collecting websites including dynamic content and rich media. It can be tailored to the needs of clients, and institutions retain full control of their collection policy (ability to select sites, specify depth, gathering frequency, etc.). The AtN platform is currently available in English and includes three modules (Administration, Collection and Report):

- The Administration Module enables the administrator to manage new and registered users.

- The Collection Module supports selection of websites and collection management (including scope, frequency, priority, access).
- The Report Module enables users to monitor reports and metrics at different levels: global archive, per collection, and per unit (captures, collected resources, size of data, and MIME type)

Quality control services can be provided on request. Most is done manually in order to meet high levels of institutional quality requirements, and IM has a dedicated QA team composed of QA assessors. IM has developed a methodology for visual comparison based on tools used for crawling and accessing data, though they are also working on improving tools and methods to deliver a higher initial crawl quality.

### Technical Solutions

IM uses several tools to complete its crawls: Heritrix v1.14.2 and v3, IM's own crawler (Memorybot) and various tailored tools for specific cases (social media, site requiring execution of pages, etc.). Most crawls are processed in an automated manner through AtN. The AtN workflow distributes crawls according to demand. Jobs are automatically sent through the workflow, supporting the automated process behind the interface that schedules, launches and makes the crawl accessible.

IM launched the development of its own Memorybot crawler for very large-scale crawls in 2010; this crawler is now operational. IM also uses its own access tool, developed in 2007, with integrated server-side link rewriting. This can run complex link rewriting rules on the fly to solve access issues such as, for example, the replacement of audio-visual files.

IM will soon move to a new infrastructure comprising a distributed storage engine that relies on HBASE and HDFS, and which automatically manages replication and fault-tolerance. This is intended to address scalability issues and improve access performance and functionality.

### Access

Access to publicly available collections is provided via the IM website. IM provides a full text search facility for most of its online collections, in addition to URL-based search. Full text search results can be integrated on a third party website and collections can be branded by owners as necessary.

Following the architecture of the Web Continuity Service by The National Archives (UK) (Spencer *et al.*, 2008), IM implemented an 'automatic redirection service' to integrate web archives with the live web user experience. When navigating on the web, users are automatically redirected to the web archive if the resource requested is no longer available online. Within the web archive, the user is pointed to the most recent crawled instance of the requested resource. Once the resource is accessed, any link on the page will send the user back to the live version of the site. This service is

considered to increase the life of a link, improves users' experience, online visibility and ranking, and reduce bounce rates.

## The Collections

Twelve web archiving collections are currently available for public browsing from the IM website, a combination of both domain and selective collections from its own and from partner institutions. One of the largest collections is the UK Government Web Archive, spanning 15 years and currently containing around 80TB of data. Other UK-based collections include a large scale crawl of local authority and National Health Service (NHS) crawls, as well as a number of local authorities' archive services. French presidential election collections are also available, as is a snapshot of the Italian domain from 2006. IM expects to receive a donation of a global crawl from IMR in 2013, which will substantially increase the size of its collection.

### 5.3. The Coca-Cola Web Archive

#### Background

The Coca-Cola Web Archive was established to capture and preserve corporate Coca-Cola websites and social media. It is part of the Coca-Cola Archive, which contains millions of both physical and digital artefacts, from papers and photographs to adverts, bottles, and promotional goods.

Coca-Cola's online presence is vast, including not only several national Coca-Cola websites but also for example, the Coca-Cola Facebook page and Twitter stream, the Coca-Cola Conversations blog, other Coca-Cola owned brands (500 in all, including for example Dr Pepper, Lilt, Oasis), and Coca-Cola brand sites (CokeZone, myCoke, Coca-Cola Heritage, Coca-Cola Store, World of Coca-Cola and many more). The first Coca-Cola website was published in 1995.

#### Collaboration

Since 2009, Coca-Cola has collaborated with Hanzo Archives and now utilizes their commercial web archiving service. Alongside the heritage benefits of the web archive, the service also provides litigation support where part or all of the website may be called upon as evidence in court and regulatory compliance for records management applications.

#### Approach

The Coca-Cola web archive is a special themed web archive that contains all corporate Coca-Cola sites and other specially selected sites associated with Coca-Cola. It is intended to be as comprehensive as possible, with integrity/functionality of captured sites of prime importance. This

includes social media and video, whether live-streamed or embedded (including Flash). Artefacts are preserved in their original form wherever possible, a fundamental principle for all objects in the Coca-Cola Archive.

Coca-Cola websites are often complex and include state-of-the-art rich media and interactive content. Efforts to capture and preserve Coca-Cola websites prior to their collaboration with Hanzo included simple flat file capture, the Internet Archive, PDF capture and Microsoft solutions. None of these proved totally sufficient for Coca-Cola's requirements.

Hanzo Archives' crawls take place quarterly and are supplemented by occasional event-based collection crawls, such as the 125th anniversary of Coca-Cola, celebrated in 2011.

### Technical Solutions

Hanzo's web archiving solution is a custom-built application. Web content is collected in its native format by the Hanzo Archives web crawler, which is deployed to the scale necessary for the task in hand. There is no limit on how many crawlers can be deployed simultaneously as the crawler is designed to be multi-machine and the machines can be deployed across numerous networks. They crawl in parallel according to the pre-determined archival policy and store resulting content in WARC files. All meaningful information, including indexed metadata and text, is extracted and stored in the Hanzo Archives Enterprise Web Archive. Multiple different representations are generated for everything captured – for example, the service generates PDF and PNG images of every page and stores them in the WARCS alongside original content. These can be delivered independently if needs be and may support later date quality assurance efforts associated with accurate rendering of the 'look and feel' of the sites.

Quality assurance is carried out with a two-hop systematic sample check of crawl contents that forces use of the upper-level navigation options and focuses on the technical shape of the site. Sites or parts of sites typically conform to a template, so at least one example of each template is checked in the QA process. Results are submitted to the team of crawl engineers to determine whether identified issues are crawl time or access-based. Patch crawls are initiated if necessary, and parameters changed for resulting crawls so that the same issue is not encountered again.

### Access

The Archive is currently accessible only to Coca-Cola employees, on a limited number of machines. Remote access is provided by Hanzo using their own access interface. Proxy-based access ensures that all content is served directly from the archive and that no 'live-site leakage' is encountered. The archive may be made publicly accessible in the future inside The World of Coca-Cola, in Atlanta, Georgia, USA.

Hanzo also provides Coca-Cola with copies of the crawl on disc, including a run time version of their access tool.

## The Collection

The Coca-Cola web archive collection is vast, containing over six million webpages and over 2TB of data. Prior to their collaboration with Hanzo, early attempts at archiving resulted in incomplete captures so early sites are not as complete as the company would like. The 1995 site exists only as a screenshot image and cannot be rendered in the same way as originally displayed, due to an issue with missing fonts. Other captures, such as from the Internet Archive or saved as PDF files, suffer from missing content or limited functionality. Sites captured by Hanzo from 2009 onwards are more complete.

The collection contains some content unique to Coca-Cola's web presence that does not exist in any other form, such as the site dedicated to their 125th anniversary. It also contains information about many national and international events for which Coca-Cola was a sponsor, including the London 2012 Olympics and Queen Elizabeth II's Diamond Jubilee. The largest sub-collections are the 'corporate' sites (e.g. the Coca-Cola company, Coca-Cola Conversations) and the Japanese collection.

## 6. Conclusions and Recommendations

Web archiving technology has significantly matured over the past decade, as has our understanding of the issues involved. Consequently we have a broad set of tools and services which enable us to archive and preserve aspects of our online cultural memory and comply with regulatory requirements for capturing and preserving online records. The work is ongoing, for as long as the Internet continues to evolve, web archiving technology must evolve to keep pace. It is encouraging to see various R&D projects developing solutions that make use of live web technology, especially to promote and facilitate access and use of the collections.

Alongside technical developments, the knowledge and experience gained through practical deployment and use of web archiving tools has led to a much better understanding of best practices in web archiving, operational strategies for embedding web archiving in an organizational context, business needs and benefits, use cases, and resourcing options. Organizations wishing to embark on a web archiving initiative must be very clear about their business needs before doing so. Business needs should be the fundamental driver behind any web archiving initiative and will significantly influence the detail of a resulting web archiving strategy and selection policy. The fact that commercial services and technologies have emerged is a sign of the maturity of web archiving as a business need, as well as a discipline.

Yet despite all of this effort, web archives still face significant challenges. The challenges of archiving social and collaborative platforms are still under investigation. Quality Assurance, ensuring that web archives harvest and archive all that they set out to, is one of the areas where least technical progress has been made over the last decade. More attention and funding are needed to develop tools that can provide, with greater reliability, validation that a crawl has not only captured all necessary files and content but also that they can be rendered appropriately. Smarter crawlers will also go some way to addressing this by increasing the reliability of a crawl and reducing the 'risk' areas currently known to exist. Without these, we run the risk of collections and archived websites that are incomplete, especially on a national domain scale. These may never fulfil their true potential and scholars will be unable to draw authoritative conclusions concerning domain level research from incomplete collections. This is particularly relevant as we begin to develop analytical access tools that encourage and support research use of very large-scale collections.

Tools to support long-term preservation of web archives are similarly under-developed. More practical R&D is needed into migration on request and emulation as specific strategies for large-scale, heterogeneous web archives. Practitioner-led research must clearly focus on business needs and deliver workable solutions in the context of holistic institutional preservation strategies. The tests carried out by the National Library of Australia in 2009 (Stawowczyk Long, 2009) provide a useful starting point, as does the web archiving work initiated in SCAPE. The growing recognition that web archives can support legal, e-discovery, and business/commercial/national intelligence services should bring significant investment into web archiving and preservation technologies in coming years.

Finally, we must not overlook the issue of legislation. Legislative challenges that not only inhibit collection but also limit access remain one of the greatest issues for collecting institutions. Until these issues are resolved, resources continue to disappear. The British Library estimates that in the absence of legal deposit legislation, less than 1% of all online activity related to the London 2012 Olympics will have been saved, and that archived websites relating to significant spontaneous UK events such as the 2005 London bombings, the 2009 UK Parliamentary expenses scandal, and the 2011 London riots are minimal, inhibiting future research (Stephens 2012). Resolving these challenges is a matter for executive bodies and politicians, but they must be informed by practical needs and requirements voiced by users and experts in the field. The technical challenges of web archiving cannot, and should not, be addressed in isolation.

## 7. Glossary

ARC	Container format for websites devised by the Internet Archive, superseded by WARC
Analytical access	A method of querying and exploring web archives at a data level rather than a page-based access experience
Archived website	One or more versions of a website in a web archive
Client-side archiving	Web archive carried out using a web client, such as a crawler
Crawl	The act of browsing the web automatically and methodically to index or download content and other data from the web. The software to do this is often called a web crawler
Crawl frequency	The regularity with which a website is crawled
Crawl logs	Files generated by web crawlers with information about the crawl
Crawler trap	The combination of issues that cause a crawler to see an excessive or infinite set of web resources, causing a crawler to become 'stuck'
Creative Commons licence	A copyright licence that allows the distribution and re-use of copyrighted works
Crown Copyright	A form of copyright used by government organizations
De-duplication	Eliminating or minimizing duplicated content
Domain name	The root of a host name, e.g. *.com, *.org, *.co.uk
Domain archiving	Archiving of an entire domain (e.g. a country code top level domain)
HTML	Hyper Text Markup Language, the main markup language for displaying pages on the Internet through a web browser
IIPC	The International Internet Preservation Consortium
Instance	A specific capture of a website in an archive, either one-time, or once in a series of captures
JHove2	A characterization tool for digital objects
OAIS	Open Archival Information System, a reference model and generic framework for building a complete archival repository
Permission	Agreement from a content owner for an archiving institution to archive their website
Petabyte	A unit of information, equates to approximately 1,000TB
QA	Quality Assurance
Scheduling	The process of specifying a crawl frequency in a web archiving system
Scope	The parameters or constraints placed on a crawl to ensure it collects only what is wanted
Seed	The starting point for a crawl, often a URL
Selective archiving	Archiving of specially selected websites on a site-by-site



	basis
Server-side archiving	An archiving solution that copies content directly from servers without using the HTTP protocol
SIP	Submission Information Package, one of the Information Packages defined in the OAIS framework
Special collection	A collection of archived websites associated with a given theme or event
Technical obsolescence	A phrase used to refer to legacy technology that is no longer easily supported or accessible on modern machines
Temporal coherence	A concept in which a digital object must represent only one given moment in time
TEP	A Title Entry Page, generated on the fly by the PANDAS web archiving system
Transactional archiving	A web archiving solution that captures all transactions between browser and server
Virus	A form of malware, designed to infect a host computer without the owner's knowledge and which may have adverse consequences
WARC	A container format for archived websites, also known as ISO 28500:2009
Web 2.0	A term coined to refer to interactive, social and collaborative web technologies, resulting in a more distinctive and more modern World Wide Web

## 8. References and Bibliography

Ainsworth, S, AlSum, A, SalahEldeen, H, Weigle, M, and Nelson, M 2013 *How much of the web is archived?* arxiv.org/abs/1212.6177 (last accessed 19-01-2013)

Archiveteam 2012 *Archiving with WGet*, Archiveteam Wiki  
[http://archiveteam.org/index.php?title=Wget\\_with\\_WARC\\_output](http://archiveteam.org/index.php?title=Wget_with_WARC_output) (last accessed 07-09-2012)

Aubry, Sara 2010 Introducing Web Archives as a New Library Service: the Experience of the National Library of France. *Liber Quarterly*, 20: 2, pp.179–199

Bailey, S & Thompson, D 2006, UKWAC: Building the UK's First Public Web Archive. *D-Lib Magazine*, 12:1 <http://www.dlib.org/dlib/january06/thompson/01thompson.html> (last accessed 14/10/2012)

Ball, A 2010, *Web Archiving*, Digital Curation Centre  
<http://www.dcc.ac.uk/sites/default/files/documents/reports/sarwa-v1.1.pdf> (last accessed 07-09-2012)

Baly, N 2006, Archiving Streaming Media on the Web, Proof of Concept and First Results, *IWAW Conference proceedings* <http://iwaw.europarchive.org/06/> (last accessed 14/10/2012)

Band, J 2012, *A new day for Website Archiving 2.0*, Association of Research Libraries  
[http://www.arl.org/bm~doc/band\\_webarchive2012.pdf](http://www.arl.org/bm~doc/band_webarchive2012.pdf) (last accessed 07-09-2012)

Binns, A 2011, *SOLR-Nutch report* IIPC <http://archive.org/~aaron/iipc/solr-nutch-report.html> (last accessed 07-09-2012)

Brindley, L 2011 *British Library international survey on E-Legal Deposit 2011: Summary of findings*, published at  
[http://www.cdnl.info/2011/pdf/e\\_2Dlegaldeposit\\_20survey\\_20CDNL\\_20Slides\\_20Aug%20\[Compatibility%20Mode\].pdf](http://www.cdnl.info/2011/pdf/e_2Dlegaldeposit_20survey_20CDNL_20Slides_20Aug%20[Compatibility%20Mode].pdf) (last accessed 07-09-2012)

British Library, 2010 *Collection Development Policy for Websites*. British Library, London.  
<http://www.bl.uk/reshelp/pdfs/modbritcdpwebsites.pdf> (last accessed 09-09-2012)

Brown, A 2006, *Archiving Websites – A practical guide for information management professionals*, 1st edn, Facet Publishing, London.

Brügger, N 2005, *Archiving Websites – General Considerations & Strategies*, The Centre for Internet Research  
[http://cfi.au.dk/fileadmin/www.cfi.au.dk/publikationer/archiving\\_underside/archiving.pdf](http://cfi.au.dk/fileadmin/www.cfi.au.dk/publikationer/archiving_underside/archiving.pdf) (last accessed 07-09-2012)

Brygfjeld, S 2002, Access to Web Archives: The Nordic Web Archive Access project, in *68th IFLA Council and General Conference Proceedings* <http://archive.ifla.org/IV/ifla68/papers/090-163e.pdf> (last accessed 14/10/2012)

Carpenter, K 2012, Net Worth. *Technology Review* online at  
<http://www.technologyreview.com/notebook/426440/net-worth/> (last accessed 07-09-2012)

Charlesworth, A 2003, *Legal issues relating to archiving of Internet resources in the UK, EU, USA & Australia*, JISC [http://www.jisc.ac.uk/uploaded\\_documents/archiving\\_legal.pdf](http://www.jisc.ac.uk/uploaded_documents/archiving_legal.pdf) (last accessed 07-09-2012)

Clarke, N 2012, *Short JHONAS presentation for the IIPC GA 2012 IIPC*  
<http://www.netpreserve.org/sites/default/files/resources/Clarke.pdf> (last accessed 21-05-2013)

COI 2008, *TG105 Archiving Websites* <http://digitalstandards.cabinetoffice.gov.uk/archiving-websites/> (last accessed 22-05-2013)

COI 2009, *TG 125 Managing URLs* <http://digitalstandards.cabinetoffice.gov.uk/managing-urls/> (last accessed 22-05-2013)

Crawford, L 2011 *Access and Analytics to the UK Web Archive* Presentation available from  
<http://www.slideshare.net/lewisdog/analytics-and-access-to-the-uk-web-archive> (last accessed 9-09-2012)

da Costa, M 2011 *Information Search in Web Archives* PhD proposal, available from  
<http://sobre.arquivo.pt/sobre-o-arquivo/information-search-in-web-archives> (last accessed 09-09-2012)

Dappert, A & Farquhar, A 2009, *Significance is in the Eye of the Stakeholder* ECDL 2009  
[http://www.planets-project.eu/docs/papers/Dappert\\_Significant\\_Characteristics\\_ECDL2009.pdf](http://www.planets-project.eu/docs/papers/Dappert_Significant_Characteristics_ECDL2009.pdf) (last accessed 07-09-2012)

Davis, M 2009, *Preserving Access: Making more informed guesses about what works*, IIPC  
<http://netpreserve.org/publications/preservingaccess.pdf> (last accessed 07-09-2012)

Day, M 2003, *Collecting and Preserving the World Wide Web*, JISC  
[http://www.jisc.ac.uk/uploaded\\_documents/archiving\\_feasibility.pdf](http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf) (last accessed 07-09-2012)

Dougherty, M et al. 2010 *Researcher Engagement with Web Archive: State of the Art*, JISC  
<http://repository.jisc.ac.uk/544/> (last accessed 09-09-2012)

DPC 2010, *Memento Project wins Digital Preservation Award* DPC  
 website <http://www.dpconline.org/newsroom/not-so-new/655-memento-project-wins-digital-preservation-award-2010> (last accessed 14-10-2012)

DPC 2012, *Web Archiving & Preservation Task Force*, Digital Preservation Coalition website  
<http://www.dpconline.org/about/working-groups-and-task-forces/524-web-archiving-and-preservation-task-force> (last accessed 14/10/2012)

Enders, M 2010 *A METS based information package for long term accessibility of web archives* iPres conference proceedings <http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/enders-70.pdf> (last accessed 07-09-2012)

Gomes, D et al. 2006, *Managing Duplicates in a Web Archive*, 21th Annual ACM Symposium on Applied Computing <http://xldb.fc.ul.pt/daniel/docs/presentations/gomes06duplicatesPPT.pdf> (last accessed 07-09-2012)

Grotke, A 2011, *Web Archiving at the Library of Congress*. *Computers in Libraries*, December 2011, available from <http://www.infotoday.com/cilmag/dec11/Grotke.shtml> (last accessed 09-09-2012)

Hanief Bhat, M 2009, *Missing Web References – A Case Study of Five Scholarly Journals*. *Liber Quarterly*, 19:2, available from <http://liber.library.uu.nl/index.php/lq/article/view/7957/8244> (last accessed 07-09-2012)

Hanzo Archives 2011, *Case Study: The Coca Cola Company*, Hanzo,  
[http://www.hanzoarchives.com/customers/the\\_coca\\_cola\\_company](http://www.hanzoarchives.com/customers/the_coca_cola_company) (last accessed 14/10/2012)

Hockx-Yu, H, Crawford, L, Coram, R and Johnson, S. 2010, *Capturing and replaying streaming media in a web archive – a British Library Case Study*, iPres conference proceedings 2010,  
<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/hockxyu-44.pdf> (last accessed 19-01-2013)

Hockx-Yu, H & Woods, T 2012, *QA Improvement in WCT*, IIPC,  
[http://netpreserve.org/events/dc\\_ga/03\\_Wednesday/WCTQAImprovement.pdf](http://netpreserve.org/events/dc_ga/03_Wednesday/WCTQAImprovement.pdf) (last accessed 07-09-2012)

IIPC 2012(a) *IIPC Future of the Web Workshop: Introduction & Overview*, IIPC  
[http://netpreserve.org/events/dc\\_ga/04\\_Thursday/Harvesting%20the%20Future%20Web/OverviewFutureWebWorkshop.pdf](http://netpreserve.org/events/dc_ga/04_Thursday/Harvesting%20the%20Future%20Web/OverviewFutureWebWorkshop.pdf) (last accessed 07-09-2012)

IIPC 2012 (b), *Preservation Working Group*, International Internet Preservation Consortium website  
<http://netpreserve.org/events/preservation-working-group> (last accessed 14/10/2012)

IIPC 2012(c) *Twitervane* <http://www.netpreserve.org/projects/twitervane> (last accessed 14/10/2012)

Internet Archive 2012, *Users of Heritrix*, IA wiki  
<https://web.archive.jira.com/wiki/display/Heritrix/Users+of+Heritrix> (last accessed 14/10/2012)

ISO, 2012, *ISO 28500:2009 Information and Documentation – the WARC file format* Available from  
[http://www.iso.org/iso/catalogue\\_detail.htm?csnumber=44717](http://www.iso.org/iso/catalogue_detail.htm?csnumber=44717)

ISO, 2013 *ISO Technical Report 14873 Information and Documentation – Statistics and quality issues for web archiving* Publication pending

ISO, 2012 *ISO 14721:2012 Space data and information systems – Open Archival Information System (OAIS) – Reference model* Available from  
[http://www.iso.org/iso/home/store/catalogue\\_ics/catalogue\\_detail\\_ics.htm?csnumber=57284](http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=57284)

Jackson, A 2012 *Formats Over Time: Exploring UK Web History*. iPres 2012 Conference proceedings, <http://arxiv.org/pdf/1210.1714v1.pdf> (last accessed 14/1-/2012)

Kahle, B 1998, *Preserving the Internet. Scientific American Special Report*, available from  
<http://web.archive.org/web/19980627072808/http://www.sciam.com/0397issue/0397kahle.html>  
 (last accessed 07-09-2012)

KEEP project website [www.keep-project.eu](http://www.keep-project.eu) (last accessed 19-01-2013)

Keskitalo, E 2010 *Web Archiving in Finland: memorandum for the members of the CDNL*, published at <http://www.cdnl.info/2010/Web%20Archiving%20in%20Finland,%20%20E-P%20Keskitalo%20-%20December%202010.pdf> (last accessed 07-09-2012)

Knight, G & Pennnock, M 2008, *Data without meaning: Establishing the significant properties of digital research* iPres Conference proceedings <http://www.significantproperties.org.uk/iPres2008-paper.pdf> (last accessed 07-09-2012)

Kulovits, H 2009 *The Planets Preservation Planning workflow and the planning tool Plato*, Planets project [http://www.planets-project.eu/docs/presentations/Kulovits\\_PLATO-wePreserveForum.pdf](http://www.planets-project.eu/docs/presentations/Kulovits_PLATO-wePreserveForum.pdf) (last accessed 07-09-2012)

Lawrence, S, Coetzee, F, Glover, E, Pennock, D, Flake, G, Nielsen, F, Krovetz, B, Kruger, A, Giles, L. 2001, *Persistence of Web References in Scientific Research*, in *Computer*, 34:2, pp.26–31 cited in Day (2003)

Leetaru, K 2012, *A Vision of the Role and Future of Web Archives*, IIPC <http://netpreserve.org/publications/Kalev2012.pdf> (last accessed 07-09-2012)

Masanés, J (Ed.) 2006, *Web Archiving*, 1st edn, Berlin: Springer

Masanés, J et al. (Eds) 2010, *International Web Archiving Workshop IAWW2012 Conference proceedings*, IAWW <http://www.iaww.net/10/IAWW2010.pdf> (last accessed 07-09-2012)

Mayr, M 2011, *IIPC Harvesting Practices Report*, IIPC [http://netpreserve.org/publications/IIPC\\_Harvesting\\_Practices\\_Report.pdf](http://netpreserve.org/publications/IIPC_Harvesting_Practices_Report.pdf) (last accessed 07-09-2012)

Mazeika, D et al. 2010, *The SOLAR System for Sharp Web Archiving LIWA/IWAW2010* <http://liwa-project.eu/images/publications/TheSOLARSystem.pdf> (last accessed 07-09-2012)

Meyer E 2010 (a), *Researcher Engagement with Web Archives: State of the Art Report*, JISC <http://ie-repository.jisc.ac.uk/544/> (last accessed 07-09-2012)

Meyer et al. 2011, *Web Archives: The Future(s)*, IIPC [http://netpreserve.org/publications/2011\\_06\\_IIPC\\_WebArchives-TheFutures.pdf](http://netpreserve.org/publications/2011_06_IIPC_WebArchives-TheFutures.pdf) (last accessed 07-09-2012)

Meyer et al. 2010 (b), *Researcher Engagement with Web Archives: Challenges & Opportunities for Investment*, JISC <http://ie-repository.jisc.ac.uk/543/> (last accessed 07-09-2012)

Mooney, P 2010, *Where have all the records gone? Challenges and Opportunities in Capturing Websites* CITRA conference presentation [http://www.citra2010oslo.no/CITRA\\_presentasjoner/Thursday/Philip\\_Mooney.ppt](http://www.citra2010oslo.no/CITRA_presentasjoner/Thursday/Philip_Mooney.ppt) (last accessed 14/10/2012)

National Archives (UK), The 2010, *Government Web Archive: Redirection Technical Guidance for Government Departments*, The National Archives

(UK) <http://www.nationalarchives.gov.uk/documents/information-management/redirection-technical-guidance-for-departments-v4.2-web-version.pdf> (last accessed 14/10/2012)

National Archives (UK), The 2011, *Tips on Finding Content in the UK Government Web Archive*, The National Archives (UK) <http://www.nationalarchives.gov.uk/documents/tips-on-finding-content-web-archive.pdf> (last accessed 07-09-2012)

National Archives (UK), The 2012, *Information on Web Archiving*, The National Archives (UK) <http://www.nationalarchives.gov.uk/webarchive/information.htm> (last accessed 14/10/2012)

NDIIPP 2012, *Science@Risk: Toward a National Strategy for Preserving Online Science*, NDIIPP <http://www.digitalpreservation.gov/meetings/documents/othermeetings/science-at-risk-NDIIPP-report-nov-2012.pdf> (last accessed 19-01-2013)

Niu, J 2012, 'Functionalities of a Web Archives', *DLib Magazine*, 18:3/4 <http://www.dlib.org/dlib/march12/niu/03niu2.html> (last accessed 07-09-2012)

Oita, M & Senellart, P 2010, Archiving Data Objects Using Web Feeds, in Masanés *et al.* (eds) 2010, available from <http://pierre.senellart.com/publications/oita2010archiving.pdf> (last accessed 07-09-2012)

OURY, C & PEYRARD, S 2011, From the World Wide Web to digital library stacks: preserving the French Web archives, in *Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES), Singapore, November 2011*, p. 237-241, available from <http://getfile3.posterous.com/getfile/files.posterous.com/temp-2012-01-02/dHqmzjcCGoexvmiBzJDCyhrhlgswoffzvsnpEAXjHFEesarwahEHrmyvj/iPRES2011.proceedings.pdf> (last accessed 21-05-2013)

Padia, K 2012, *Visualising Archive Collections at Archive-It*, MSc thesis <http://www.cs.odu.edu/~mweigle/papers/padia-thesis12.pdf> (last accessed 07-09-2012)

Pennock, M 2011, Twitterlane: Crowdsourcing Selection, on *UK Web Archive Blog* (Dec 2011) <http://britishlibrary.typepad.co.uk/webarchive/2011/12/twitterlane.html> (last accessed 09-09-2012)

Pennock M & Davis R, 2009, *ArchivePress: A Really Simple Solution to Archiving Blog Content*, iPres Conference proceedings <http://escholarship.org/uc/item/7zs156mb> (last accessed 07-09-2012)

Pinsent, E 2009, *Working with the Web Curator Tool (part 2): wikis*, blog post on ULCC's Da Blog <http://dablog.ulcc.ac.uk/2009/03/10/working-with-web-curator-tool-part-2-wikis/> (last accessed 07-09-2012)

Pop, R *et al.*, 2010, *Archiving Web Video*, LiWA project publication <http://liwa-project.eu/images/publications/ArchivingWebVideo.pdf> (last accessed 19-01-2013)

Potter, A 2010, *Web Archiving & History of IIPC*, IIPC [http://netpreserve.org/events/2010GApresentations/01\\_Tutorial\\_History%20of%20Web%20Archiving\\_N\\_IIPC.pdf](http://netpreserve.org/events/2010GApresentations/01_Tutorial_History%20of%20Web%20Archiving_N_IIPC.pdf) (last accessed 07-09-2012)

Raditsch, M 2012 *Web Archive Mime-Type detection workflow based on Droid and Apache Tika* SCAPE wiki, <http://wiki.opf-labs.org/display/SP/SO17+Web+Archive+Mime-Type+detection+workflow+based+on+Droid+and+Apache+Tika> (last accessed 07-09-2012)

Raditsch, M *et al.* 2012, *Web content executable workflows for experimental execution*, SCAPE project [http://www.scape-project.eu/wp-content/uploads/2012/05/SCAPE\\_D15.1\\_ONB\\_v1.0.pdf](http://www.scape-project.eu/wp-content/uploads/2012/05/SCAPE_D15.1_ONB_v1.0.pdf) (last accessed 07-09-2012)

Rynning, M *et al.*, 2011, *BlogForever: D2.4 Weblog spider prototype and associated methodology*, BlogForever project [http://blogforever.eu/wp-content/uploads/2011/11/BlogForever\\_D2\\_4\\_WeblogSpiderPrototypeAndAssociatedMethodology.pdf](http://blogforever.eu/wp-content/uploads/2011/11/BlogForever_D2_4_WeblogSpiderPrototypeAndAssociatedMethodology.pdf) (last accessed 07-09-2012)

Sorensen *et al.* 2012 *NetarchiveSuite: A complete totolset for webarchiving at both large and small scales* IIPC [http://netpreserve.org/events/dc\\_ga/04\\_Thursday/Netarchive/nas.ppt](http://netpreserve.org/events/dc_ga/04_Thursday/Netarchive/nas.ppt) (last accessed 07-09-2012)

Spencer *et al.* 2008, *UK Government Web Continuity: Persisting Access through Aligning Infrastructures*, IJDC 1:4, 2009 [www.ijdc.net/index.php/ijdc/issue/view/7](http://www.ijdc.net/index.php/ijdc/issue/view/7) (last accessed 19-01-2013)

Stack, M 2005 *Full Text Search of Web Archive Collections* IAWAW2005 <http://iawaw.europarchive.org/05/stack3.pdf> (last accessed 07-09-2012)

Stawowczyk Long, A 2009, *Long term Preservation of Web Archives – Experimenting with Emulation and Migration Technologies*, IIPC [http://netpreserve.org/publications/NLA\\_2009\\_IIPC\\_Report.pdf](http://netpreserve.org/publications/NLA_2009_IIPC_Report.pdf) (last accessed 07-09-2012)



Stephens, A 2012, *Towards an 'advocacy pack' for e-legal deposit*, Conference of Directors of National Libraries [http://www.cdnl.info/2012/pdf/Annual%20Meeting/Andy\\_Stephens.DOC](http://www.cdnl.info/2012/pdf/Annual%20Meeting/Andy_Stephens.DOC) (last accessed 05-09-2012)

Thompson, D 2008, Archiving websites, in *DCC Curation Manual* <http://www.dcc.ac.uk/sites/default/files/documents/resource/curation-manual/chapters/archiving-web-resources/archiving-web-resources.pdf> (last accessed 07-09-2012)

ULCC, University of London Computer Centre and UKOLN 2008, *Preservation of Web Resources Handbook*, JISC <http://www.jisc.ac.uk/publications/programmerelated/2008/powrhandbook.aspx> (last accessed 07-09-2012)

Van de Sompel, H 2012 *Release of SiteStory Transactional Archive Solution* on MementoWeb website (Aug 2012) <http://www.mementoweb.org/news/node/40> (last accessed 09-09-2012)

Van de Sompel *et al.*, 2009 *Memento, Time Travel for the Web* <http://arxiv.org/abs/0911.1112> (last accessed 14-10-2012)

Webcite 2012, *About WebCite*, Webcite website <http://www.webcitation.org/faq> (last accessed 14/10/2012)

Weiss, R 2003, *On the Web, Research Work Proves Ephemeral*, Washington Post Nov 23rd 2003, available from [http://faculty.missouri.edu/~glaserr/205f03/Article\\_WebPub.html](http://faculty.missouri.edu/~glaserr/205f03/Article_WebPub.html)

### Websites & downloads

Archive-It: <http://www.archive-it.org/>

Archivethe.Net: <http://archivethe.net/en/>

DataCite: <http://www.doi.org/>

Diigo: <http://www.diigo.com/>

HanzoArchives: <http://www.hanzoarchives.com/>

HTTrack: <http://www.httrack.com/>

Heritrix: <https://webarchive.jira.com/wiki/display/Heritrix/Heritrix#Heritrix-Downloads>

International DOI Foundation: <http://www.doi.org/>

International Internet Preservation Consortium: <http://netpreserve.org/about/index.php>

Internet Memory Foundation & Internet Memory Research: <http://internetmemory.org/en/>

Java Web Archive Toolkit: <https://sbforge.org/display/JWAT/Overview>

JhoNAS: <http://netpreserve.org/projects/jhonas>

MEMENTO: <http://www.mementoweb.org/>

Mendeley: <http://www.mendeley.com/>

NetarchiveSuite: <https://sbforge.org/display/NAS/NetarchiveSuite>

NutchWax: <http://sourceforge.net/projects/archive-access/files/nutchwax/>

PANDAS: <http://pandora.nla.gov.au/pandas.html>

ReedArchives: <http://www.reedarchives.com/>

SiteStory: <http://mementoweb.github.com/SiteStory/>

SOLR: <http://lucene.apache.org/solr/>

Twittervane: <https://github.com/ukwa/twittervane>

Uc3 Web Archiving Service: <http://www.cdlib.org/services/uc3/was.html>

WARC Tools Project: <http://netpreserve.org/projects/warc-tools-project>

Wayback: <http://sourceforge.net/projects/archive-access/files/wayback/>

Web Curator Tool: <http://webcurator.sourceforge.net/>

WGet: <http://www.gnu.org/software/wget/>

[Workshop on quality indicators:](#)

[http://www.netpreserve.org/sites/default/files/resources/IIPCGA\\_ISO\\_Workshop.pdf](http://www.netpreserve.org/sites/default/files/resources/IIPCGA_ISO_Workshop.pdf)

Zotero: <http://www.zotero.org/>