

The Past Issue of the Web

Helen Hockx-Yu
The British Library
St Pancras, 96 Euston Road
London, NW1 2DB
United Kingdom
+44 (0)20 7412 7184
Helen.hockx-yu@bl.uk

ABSTRACT

This paper takes a critical look at the efforts since the mid-1990s in archiving and preserving websites by memory institutions around the world. It contains an overview of the approaches and practices to date, and a discussion of the various technical, curatorial and legal issues related to web archiving. It also looks at a number of current projects which take a different approach to dealing with the temporal aspects or persistence of the web. The paper argues for closer collaboration with the main stream web science research community and the use of technology developed for the live web, such as visualisation and data analytics, to advance the web archiving agenda.

Categories and Subject Descriptors

H.3 [INFORMATION STORAGE AND RETRIEVAL]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.5 Online Information Services; H.3.6 Library Automation; H.3.7. Digital Libraries. H.5 [INFORMATION INTERFACES AND PRESENTATION]. K.4 [COMPUTERS AND SOCIETY]. K.5 [LEGAL ASPECTS OF COMPUTING].

General Terms

Management, Design, Economics, Reliability, Human Factors, Standardization, Legal Aspects.

Keywords

Heritage, Academic Research and the Web, Web Archiving, Library Information Management, Digital Libraries, Web Archive, Web Harvesting, Electronic Legal Deposit, Digital Preservation.

1. INTRODUCTION

The World Wide Web is an information system which has witnessed unprecedented growth in the last 20 years since its birth in 1991. It plays an undisputed important role in modern society, fundamentally changing the way we live and communicate. Its impact has been felt in how we publish, learn, teach and research, and many other areas of human activities. The current and

transient nature of the Web means that new information replaces older information constantly without any records of the previous state (or versions) of the same information. While new information is being added, existing information also disappears from the web, leave a significant gap in our knowledge of the historical web and potentially in social history.¹ It is therefore not surprising that memory institutions around the world quickly realised the need and value of collecting the content on the Web and started the epic journey of archiving and preserving it since the mid-1990s, as “An archive of the Internet may prove to be a vital record for historians, businesses and governments.”[20] The Internet Archive’s Wayback Machine [15] is the earliest and most comprehensive web archive to date, containing over 150 billion web pages archived from 1996. National libraries and archives, which traditionally have the duty to preserve a nation’s cultural and scientific heritage, also actively archive culturally important websites. Many of them are members of the International Internet Preservation Consortium (IIPC) [12]. A registry of the members’ web archives provides an overview of what has been archived in terms of geographical and temporal coverage [13].

2. KEY APPROACHES AND PROCESSES

2.1 Domain versus Selective Archiving

Web Archiving refers to the activities of selecting, capturing, storing, preserving and managing access to snapshots of websites over time. Determined by the strategic importance perceived by the archiving institution, resource available and sometimes legal requirements, diverse approaches have been taken to archive content on the web, ranging from capturing individual web pages to entire top-level domains. Dependent of the scale and purpose of collection, however, a distinction can be made between two broad categories: domain archiving and selective archiving.

Domain archiving is intended to capture a snapshot of the state of an entire domain (or a subset such as a national domain) at a given point in time, resulting in large scale web archive collections. The best known domain archive is the previously mentioned Internet Archive’s Wayback machine, which was established with the goal to preserve the global web. Domain harvesting is a fairly automated process but limited by the scale of operation, it tends to take only a shallow scoop of the top level pages of a website, lacking the completeness or depth required by some researchers.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WebSci '11, June 14-17, 2011, Koblenz, Germany.
Copyright 2011 ACM.

¹ Research done by the British Library based on discovery crawls of approximately 4.5 million distinct UK domains indicates that approximately 5% of the domains have become inactive between June and December 2010.

Selective archiving is performed at much smaller scale, more focused and undertaken more frequently. A selection process takes place to identify relevant websites based on criteria such as theme, event, significance or relevance. The British Library, for example, has been selectively archiving UK websites since 2004 and the selection process is driven by a formal Collection Development Policy, prioritising the inclusion of websites covering the following broad categories:

- reflect the diversity of lives, interests and activities throughout the UK
- contain research value or are of research interest
- feature political, cultural, social and economic events of national interest
- demonstrate innovative use of the web [3]

Quality assurance, the evaluation of harvested websites to determine whether pre-defined quality standards are being attained, is a common element of selective archiving. This currently heavily relies on visual comparison, review of previous harvests and crawl logs. A selective web archive also tends to have more descriptive metadata, often added by curators during the selection or after the harvesting process, which can be used to build richer search and browsing functions in the User Interface of a web archive. The UK Web Archive provided by the British Library is a good example. [30]

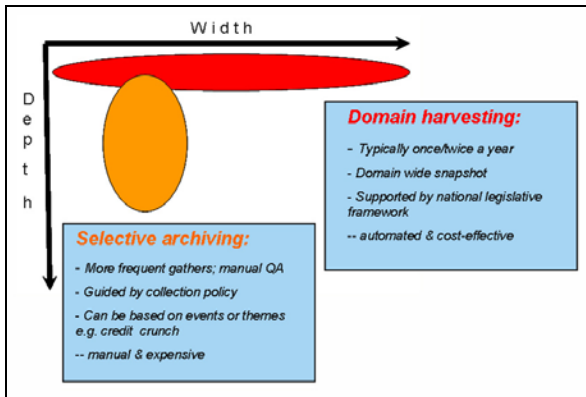


Figure 1. Domain versus selective archiving

Some websites update frequently and the changes will not be captured by just relying on infrequent annual domain harvests. It is not uncommon for a single archiving organisation to take a combination of both approaches, maximising the comprehensiveness and replicability of key websites through additional and or deeper harvests.

With the exception of the Internet Archive and the European Archive [9], large scale domain harvesting is commonly performed by national libraries and archives, enabled by national legislations [33]. One of these is the Legal Deposit legislation. Governed either by Copyright law or separate legislations, Legal Deposit for printed publications has a long established history in many countries. For example, it has existed in English law since 1662. It is a statutory obligation which requires publishers to deposit one or more copies with a designated national institution (e.g. national library, parliamentary library, national archives or main government or university library). Its purpose is to collect the nation's published output systematically and as

comprehensively as possible, for use by current and future generations.

In the past decade, many countries have been reviewing or modifying their legislations to address the challenges of electronic publications. In countries such as France, Denmark, Norway and Austria, Legal Deposit legislations have been revised to include websites, allowing national libraries and archives to harvest websites at national domain level without having to ask for permission from website publishers. In the UK, the Legal Deposit Libraries Act 2003 introduced a framework for the deposit of non-print works, which requires further legislation to be introduced to define the procedures [7]. A commonality of web archives collected for the purpose of Legal Deposit is the restrictive access requirement, often limited to on-site access only.

University libraries are also key participants in web archiving. Instead of building large-scale, multi-purpose web archives like the national institutions, university libraries often develop web archives which are research-led and with disciplinary focuses. A common model is for the libraries to offer web archiving as a service to academic departments, while being responsible for operating the infrastructure and developing the necessary tools to enable web archiving. Their web archive collections tend to have more involvement from researchers both in terms of selection and use. Both Harvard University Library [10] and the California Digital Library [5] take this distributed approach to selective archiving.

Researchers cannot always find useful or relevant material in large multi-purpose archives built by librarians and archivists or are simply not aware of their existence. So they develop their own archives using desktop archiving tools such as Offline Explorer and save copies of websites on local disks for use in their research. These personal archives are purpose-built, highly selective and specific to research topics or projects. They are however not available to others and there is little thought with regard to longevity.

2.2 Key Processes of Web Archiving

Regardless of the approach, there is a set of processes essential to web archiving which need to be performed and managed to ensure fitness for purpose of any web archiving system. This framework may be a simplification but it provides a high-level overview of the wide range of complex tasks the archiving organisations have to perform. In addition, it helps us to critique or analyse later in the paper the current practices related to these processes.



Figure 2. Key processes of web archiving

Selection is the decision-making process which determines what websites to archive and to include as part of a web archive collection.² Descriptive metadata may be added to describe the selected content. For institutions which do not have the legal mandate to collect websites, seeking permission is also part of the workflow.

Harvesting (or crawling) refers to the automated process of downloading copies of selected websites, commonly using web crawling software. It generally starts from a list of URLs (seeds), visiting and downloading them, before identifying all the hyperlinks within the visited pages and recursively visiting and downloading these too. Quality assurance checking, manual or automated, often forms part of the harvesting process.

Storage refers to the process of retaining archived websites on a storage medium securely and reliably. Commonly used archival formats for websites are ARC [4] and WARC [18, 19]. Both are container formats developed specifically for web archives.

Access refers to replaying and providing access to the archived websites for the defined users.

Digital Preservation refers to the standards, best-practices and technologies which together are needed to ensure access to web archives over time.

3. THE WEB ARCHIVING PARADOXES

There are many paradoxes embedded in the environment in which web archiving operates. They concern complex management, curatorial, legal and technical issues and act as contradictory forces which make it very difficult for archiving institutions to decide where to focus energy and resource.

In the fourteen years since the attempt to keep a record of the live web began, progress no doubt has been made and there is in general increased awareness and acceptance of web archiving. Thanks to those involved in web archiving, a snapshot has been kept of a significant portion of the web. A global community of practitioners has gradually been formed, mainly around the IIPC, where there is a surprisingly high-level of uniformity in the use of technology. The recent survey of web archiving initiatives conducted by the Portuguese Web Archive shows a high concentration in terms of the use of technology, mainly around Heritrix, NutchWax and Wayback.³ This a set of open-source tools developed by the Internet Archive which can be used to crawl websites, produce full-text indexes and replay archived websites. One can think of many advantages: interoperability, development of standards, transferable lessons and practices, and collaboration. Equally there are disadvantages associated with closed community, in that it is easy to accept certain ways of doing things and become less responsive to changes happening

² Even in the context of broad domain harvesting, selection is still applicable. This includes assembling seeds list to start a crawl, filtering out spam and websites with crawl traps etc. For national libraries collecting web content as Legal Deposit material, there are additional requirements to select content based on territoriality criteria. An important difference is that this type of selection quite often can be automated and does not solely involve curatorial judgement.

³ A rough count shows that nearly 70% of the 48 web archiving initiatives use at least one of the three main tools [33].

elsewhere beyond the community. It can also be argued, from a risk management perspective, that it is not a good practice to be dependent on any single approach and not to have more options.

3.1 The Need for Archiving the Web

The benefits of the web archiving community currently outweigh significantly the limitation and risk mentioned above, so that they do not make the top of the list when it comes to the fundamental issues related to web archiving. Archiving institutions often find themselves in a situation where they have to explain why the web needs to be archived, or are challenged by the relatively low use of the web archives when they are asked to justify the spending. Masanes gives a good account of the arguments against web archiving:

- The quality of the information on the web is not good enough for long-term preservation.
- The web is a self-preserving medium. Based on the rule of survival of the fittest, information which deserves to be preserved will remain on servers and the rest will disappear.
- Archiving the web is an impossible task, due to its size, privacy concerns and various rights issues. [24]

In addition to the challenges and skepticism from the areas of library, publishing and computing science, a more fundamental question comes from the ordinary citizens who ultimately will be footing the bill for electronic Legal Deposit: the web is full of rubbish and why should tax payers money be spent to collect it? Libraries in modern times have not been challenged to defend the principle of Legal Deposit, nor to justify their Legal Deposit operations by usage statistics of the deposited print publications. However, when Legal Deposit was first implemented in 1537 by King François I of France, the decree was not well respected. It was even abolished during the French Revolution in the name of liberty, only to be reinstated in 1793 [12]. It has taken a few hundred years for Legal Deposit to become an established principle and to be considered as “the foundation of a national policy of freedom of expression and access to information.” [12] In this process, legal text has also changed to meet new requirements and accommodate new types of publications. The web is only 20 years old and web archiving has an even shorter history. It will take some time for the web to be embedded as part of the Legal Deposit framework and for the value of web archives to be fully realised and utilised.

3.2 The Role of the Archivist and Curator

Archivists, librarians and curators are closely related professions who have traditionally acted as “gatekeepers” of information. It is not in the scope of this paper to discuss the differences between these professions,⁴ and the distinctions have blurred anyway in

⁴ The Society of American Archivists defines the archives profession and explains the conventional differences between these professions: “The work of the archivist is related to, but distinct from, that of certain other professionals. The librarian and the archivist, for example, both collect, preserve, and make accessible materials for research; but significant differences exist in the way these materials are arranged, described, and used. The records manager and the archivist are also closely allied; however, the records manager controls vast quantities of institutional records, most of which will eventually be

the context of digital information. When it comes to the web, the complex and multi-dimensional space where many communications and interactions take place, it is not always possible or helpful to map the existing concepts or framework. It is sufficient to understand that all these professions are responsible for collecting and organising information and help people to find and use it.

A key task performed by archivists and curators in web archiving is selection, making decisions about what websites to include in a web archive. This is not a straightforward task: selecting something at the same time means excluding something else. This requires good knowledge of the collections' content as a whole, as well as an overview of the pool of information to select from. In addition, applying selection criteria, often too broad and meant to be guidance, to individual websites seems hardly an objective process. One can easily challenge why an individual website is or is not selected. The analogy of trying to find a needle in a haystack has often been used to describe the frustration related to selection for web archives.

The focus on collection development is rooted in the view that the web consists of historical documents (web pages) which are mainly used for reference. The curator's role is to filter or select them, label or describe them, arrange or group them together in the same way as printed books and journals. This view dominates the current web archiving practices in many ways, for example in defining and assessing quality where the high end of the spectrum requires faithful replica of websites, and in the design of user interfaces for web archives where the main way of navigation is browsing individual websites page by page. This simply does not scale up to the level of the global web or a national domain. More importantly, the static "document-centric" view of the web does not seem to do it justice as it could miss out on the exciting possibilities the web offers scholars, for example as a corpus of aggregated data, which in addition contain context, relationships and many embedded not so obvious trends and pattern.

In a fast evolving, non-hierarchical, highly open and participatory environment such as the web, there is no longer a single or even authoritative view on the same things. The paradox is that while the web offers a wealth of cultural, social and scientific information, it also challenges the role of the professionals who traditionally safeguard and preserve this information. Heritage institutions need to think how best to fulfill their canonical mission in a changed and changing world. Masanes observes similar issues but remains positive when discussing the role of librarians in web archiving: "...But the fact that manual selection of content does not scale to the Web size is not a reason for rejecting web archiving in general. It is just a good reason to reconsider the issue of selection and quality in this environment." [24]

destroyed, while the archivist is concerned with relatively small quantities of records deemed important enough to be retained for an extended period. The museum curator and the archivist are associated; however, the museum curator collects, studies, and interprets mostly three-dimensional objects, while the archivist works with paper, film, and electronic records. Finally, the archivist and the historian have had a longstanding relationship; the archivist identifies, preserves, and makes the records accessible for use, while the historian uses archival records for research." [27]

3.3 The Technical Challenges

The skepticism about web archiving is closely related to the technical challenges currently facing the community, the most fundamental being the rapidly changing web with new formats, protocols and platforms, requiring the archiving organizations to respond to its continuous development and improve the capability to archive new content types as they emerge.

In 2010, Ball compiled a comprehensive list of tools employed in web archiving, developed to support many of the key processes described in 2.2, including crawlers, workflow management tools, archival formats, indexing, processing and rendering tools. This is the most up-to-date overview of the portfolio of web archiving technology [2]. Deploying these would technically enable an organisation to undertake domain or selective archiving, doing a good job capturing and replaying the static portion of the web and content which can be served by requesting a URL. However, as discussed below, the archiving technology is still not adequate to deal with the web in full, leaving certain types of content on the web out of reach.

Harvesting is straightforward when URLs can be determined. The crawler will be able to download a copy of the file via a simple HTTP request, by going to the right URL. When the URLs are not explicitly referenced in HTML but embedded in JavaScript or Flash presentations or generated dynamically based on the results of some interactions with the users, the crawlers are not able to capture the content, bringing back only the static HTML elements of a website.

The recursive URL-based crawling method falls short of collecting systematically an increasingly bigger portion of the web, including content behind web forms and query interfaces, commonly known as the "deep web", streaming media, content delivered over non-http protocols and social media. A recent addition added to the list of challenges is the semantic web and linked data [8].

A commonly cited technical issue related to large scale crawling is temporal incoherence. This happens when the rate of capture is exceeded by the rate with which the websites are being updated or refreshed, resulting in a distorted snapshot with the co-existence of web pages with different lifespan. The LiWA project has researched this topic and developed a tool to visualise and help identify coherence defects [22].

Capturing the content only completes half the job. In order to provide access, archived websites need to be replayed to the end users, which is equally a challenge. The current tools, for example, do not have the capability of playing back streaming media content embedded in archived websites. Both the British Library and the European Archive have made progress in archiving web videos.⁵

A common problem in replaying archived content is the so-called "live leakage", which occurs when links in an archived webpage

⁵ Both the British Library and the European Archive have made progress in capturing and replaying streaming media content on the web. The former successfully developed a solution to capture and replay a large scale public arts project's website containing over 2,400 hours of Flash videos [11]. The latter used the Rich Media Capture module as a Heritrix external plug-in, developed by the LiWA project, to capture web video [25].

resolve to the current copy on the live site, instead of pointing to the archival version within the web archive. This is usually caused by incorrect URL-rewrite, often a result of links embedded in JavaScript not being detected by Wayback, the commonly used replay tool for web archives. Wayback does have a so-called “proxy mode”, which allows the browser to be configured to proxy all HTTP requests through the Wayback Machine [16]. This stops the leakage to the live web and helps identify gaps in the web archive but can be impractical to implement for publicly accessible archives as this may conflict with existing proxy infrastructure, preventing users from configuring the proxy setting of the browser [17].

3.4 The Access Problem

Many web archives do not have open access, especially those developed with a Legal Deposit mandate. Those providing open access either seek permission from the IPR holders, which often involves high costs, or decide to take and manage a certain level of risks.

While most web archives struggle to encourage usage and understand researchers’ needs, the restrictive access to web archives has been seen by some as another paradox for web archiving, especially when this concerns information which is or has been openly available on the web.⁶ For most archiving organisations, compiling use cases is a common way to understand user requirements and help develop interfaces to the web archive. The problem is that many of these use cases are semi-hypothetical without real users behind to support them or verify the need [14]. It is not until recently that cases of close contact with researchers have begun to emerge. An example is the British Library’s Researchers and the UK Web Archives Project, involving eight researchers from various disciplines in the humanities, who will work with curators to develop research-led collections and to provide feedback on web archive functionalities [27]. The experience of the project so far tells us that there are many commonalities among the researchers but also much differences in research methods, use of resource and content expected in a web archive. In other words, they all want different things. Another observation is that perhaps the web archiving organization should take the lead in developing and demonstrating new access methods for web archives, showing what is possible, so that the researchers can tell us what is useful. We all know about the “chicken and egg” scenario where it is difficult to articulate requirements based on abstract concepts.

Thomas et al. argue that for web archives to find value in the research world, multiple access points are required: administrative, descriptive and contextual [8]. These may mean different things to different researchers and there may be many ways to meet such requirements but the significance of this recommendation is that it takes us beyond the current way of accessing web archives. Instead of using them for reference, web archives contain aggregated datasets that can also be used for analytics.

3.5 The Legal Issues

The legal issues reported by the JISC and the Wellcome Trust in the 2003 study of “Legal issues relating to the archiving of Internet resources in the UK, EU, USA and Australia”, are still

⁶ The legal reason for this is explained later in 3.5.

highly applicable in today’s web archiving environment [7]. Archiving websites without permissions breaches the copyright law. Providing access to archived websites, even with permission, could be regarded as republishing and thus transfers certain legal risks, such as libel, from the original publishers. With domain-wide automatic harvesting, there is the additional risk of accidentally collecting illegal and undesirable content such as pornography and terrorism-related material.

Some libraries have chosen a permission-based approach to minimise the legal risk by contacting website owners prior to archiving. This is a fairly manual and costly process, resulting in highly selective and patchy web archive collections, as in practice the common rate of success with permissions is between 25% to 35% percent. A paradox is that website owners are often not in a position of granting the libraries permissions, even if they are supportive of web archiving, due to third party rights issues.

A key development in the UK is the move towards electronic Legal Deposit. The government has confirmed their commitment to deliver regulations for electronic Legal Deposit, which will include online content that can be obtained through a harvesting process [7].

It is important to realise that Legal Deposit does not remove the legal issues related to web archiving. This is very well explained by the UNESCO guidelines for Legal Deposit Legislations: “There are two major problems related to legal deposit of electronic or digital material vis-à-vis copyright. The first is related to the deposit process itself. The legal deposit of electronic publications necessitates the reproduction of protected works...Since digital material might have to be collected through downloading from the master copy on a server, the process raises the issue of permission to reproduce a protected work. Again, national copyright legislation or legal deposit legislation should provide legislative permission to reproduce documents for legal deposit purpose.”

The second issue to deal with is access. Considering that it is widely recognized, at both the national and international level, that a copyright owner has an exclusive right to communicate a protected work to the public and that most electronic publications need to be “communicated to the public” in order to be seen and read, the deposit copy of such electronic publications might require a specific exception allowing access to the clientele of the national legal deposit institution” [21].

4. THE WAY FORWARD

This paper is intended to reflect the progress in web archiving and explore a way forward which copes better with the evolving web and research needs. The earlier discussion of the paradoxes should not overshadow the significant achievement by the community. There is now a range of tools available, supporting key web archiving processes. More and more countries have set up web archiving programmes and the IIPC membership has extended from the initial six to forty. It is not just the traditional heritage institutions that are involved in preserving the web. Universities and researchers are also taking part in this effort and commercial archiving services have started to appear.

4.1 New Developments

There are a number of new developments which need to be taken into account when considering future directions for web archiving. They indicate trends and provide inspiration for new

ways of thinking which will help archiving organisations to develop and expand into the future in response to the evolving web. Some of these initiatives also promise to deliver concrete and practical solutions which can be utilised by many archiving organisations.

4.1.1 JISC Studies on Web Archives

Two agenda-setting studies on web archiving were published in 2010, both funded by the Joint Information Systems Committee (JISC) and carried out by the Oxford Internet Institute, University of Oxford. Titled *Researcher Engagement with Web Archives, State of the Art*, the first report is a comprehensive survey of the state of the art in web archiving, with a focus on its relationship with individual researchers and their research needs. It concludes that to reach the potential of web archives as objects of research, it needs to be taken more seriously as an important element of research programs involving web resources [8].

A companion study titled *Researcher Engagement with Web Archives, Challenges and Opportunities for Investment* looks beyond the current state of the art and points out some important opportunities for investment, which could move web archiving technology and practices to the next level of comprehensiveness and usefulness. The most significant recommendation of all, however is the following: "A move away from costly and time-consuming attempts to identify a priori the content (the "needles") likely to be of interest to web researchers, and towards what we call "collecting the haystacks": the rapidly-falling cost of storage, and new technologies and metadata conventions for managing multi-petabyte repositories, suggest that less effort should be placed on selection and collection strategies, and more on ways for users rapidly to survey, annotate, contextualise, and visualise those repositories, and to find and select the thematic elements of interest to them." [29]

The report also recommends closer integration with the Web Science community, which is already researching many of the issues with which the web archiving community is currently struggling.

4.1.2 The Memento Project

The Memento project [25] is a collaboration project between researchers at the Los Alamos National Laboratory and the Old Dominion University.

Memento proposes a protocol-based framework which adds temporal dimension to the HTTP protocol so that archived versions (called *Mementos*) of a resource can be serviced seamlessly by the web server which holds the original resource. This is achieved by qualifying the HTTP request with a *DateTime* parameter. If web servers can honour the request, it will simply serve the page. In case it does not hold the memento of that resource, it redirects to a server that does. This can be a web archive that has the best archival coverage of the requested resource or an aggregator, which holds metadata from various web archives and has the ability of redirecting a client to a *memento* in response to a specific date / time.

Memento deals with the temporal aspect of the web at the protocol level. While tapping into web archives in machine to machine manner, it offers end users seamless access to the past version of a resource by specifying date/time in a browser plug-in, without having to leave the current browsing environment or visiting a specific web archive interface. Memento is currently

planning a joint project with a number of member institutions of the IIPC to aggregate and make discoverable metadata from web archives.

4.1.3 Zotero

Zotero [34] is an open source reference management tool which has been found useful by many researchers. This is another tool which has machine access into a web archive. It is integrated with parts of the Internet Archive's existing collections which allow researchers to select already archived files and web pages from the Internet Archive's existing collections and add these to their on-line library [35].

4.1.4 Longitudinal Analytics of Web Archive Data (LAWA)

The LAWA project [23] is a three-year project funded by the European Commission to develop infra-structure, methods, and software tools for aggregating, querying, and analyzing large scale web archive data. The project has a particular focus on data analysis along the time dimension for web data that has been crawled over time.

4.1.5 Archive Communities Memories (ARCOMEM)

Another important initiative funded by the European Commission is the ARCOMEM project [1], which is about developing solutions to help memory institutions exploit the social web. The intended outcomes of the project include among others:

- Innovative models and tools for Social Web driven content appraisal and selection, and intelligent content acquisition
- Novel methods for Social Web analysis, Web crawling and mining, event and topic detection and consolidation, and multimedia content mining.

4.2 The Value of the Haystacks

A traditional view is that researchers access previous states of individual web pages and sites in a web archive. The above-mentioned new developments demonstrate a shift of focus in web archiving, from human access to machine access and from the level of single webpages or websites to the entire web archive collection. There is a realisation that there may be significant value in the "haystacks". Using visualisation and data analytic techniques, there are opportunities to provide access to different views of a web archive, unlocking embedded patterns and trends, relationships and contexts. Taking this thought further and accepting that it is almost impossible to capture a 100% faithful representation of the web, perhaps it is not so disastrous if a few pages out of tens of millions are not captured fully due to limitations of the crawler?

Some preliminary work has been undertaken by the British Library to develop data-based visual tools to access the UK Web Archive's content as alternatives to the standard search and browse functions. Tag clouds have been generated by analyzing the Special Collection "UK General Election 2005", which consists of web pages with electoral content archived during and immediately after the UK general election campaign of 2005 [32]. The Archive also has a visual browsing tool, incorporating CoolIris [6], offering navigation through multiple thumbnail images of archived web pages [31]. The British Library intends to continue the development in analytical access, focusing on the entirety of the web archive data. It would be useful if more web

archives would explore this, eventually allowing comparisons of various archives.



Figure 3. 3D wall for visual browsing in the UK Web Archive

5. CONCLUSIONS

The past issue of the web has not passed. There is plenty of scope for further development. The web archiving community should really look beyond the current practices and take advantage of the many powerful technologies designed for the live web.

It is important that the momentum already gathered behind web archiving is not lost, especially at a time when resources are being severely constrained for many memory institutions.

A problem for organizations undertaking large scale web archiving is the array of issues they have to deal with and the constant pressure to keep pace with the evolving web. Often tasked with running operations and services, memory institutions are not best placed to carry out dedicated research and development addressing the issues they face.

Web Science as a new interdisciplinary field seems a natural home for web archiving. Not only does the discipline study many of the aspects of the object web archives collect, it may also make use of web archives to understand the past and evolution of the web.

Consider this a plea for help.

6. ACKNOWLEDGMENTS

Thanks from the author go to Lewis Crawford, Web Archiving Lead, the British Library, and Professor Michel Hockx, School of Oriental and African Studies, University of London, for stimulating discussions, verification of technical and scholarly details, as well as encouragement and editing of the paper.

7. REFERENCES

- [1] Archive Communities Memories (ARCOMEM). <http://www.arcomem.eu/>.
- [2] Ball, A. 2010. *Web Archiving*. Digital Curation Centre, Edinburgh.
- [3] British Library Web Archiving Programme. 2010. The British Library collection development policy for websites. <http://www.bl.uk/reshelp/pdfs/modbritcdpwebsites.pdf>.
- [4] Burner, M. and Kahle, B. 1996. ARC file format. <http://www.archive.org/web/researcher/ArcFileFormat.php>.
- [5] California Digital Library Web Archiving Service. <http://www.cdlib.org/services/uc3/was.html>.

- [6] CoolIris. <http://www.cooliris.com/>.
- [7] Department for Culture, Media and Sport. Legal deposit. http://www.culture.gov.uk/what_we_do/libraries/3409.aspx.
- [8] Dougherty, M., Meyer, E.T., Madsen, C., Van den Heuvel, C., Thomas, A. and Wyatt, S. 2010. *Researcher Engagement with Web Archives: State of the Art*. JISC, London.
- [9] European Archive. <http://www.europarchive.org/>.
- [10] Harvard University Library Web Archive Collection Service. <http://wax.lib.harvard.edu/collections/home.do>.
- [11] Hockx-Yu, H., Crawford, L., Coram, R. and Johnson, S. Capturing and replaying streaming media in a web archive—A British Library case study. <http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/hockxyu-44.pdf>.
- [12] International Internet Preservation Consortium (IIPC). <http://netpreserve.org/about/index.php>.
- [13] IIPC. Member archives. <http://netpreserve.org/about/archiveList.php>.
- [14] IIPC Access Working Group. 2006. Use cases for access to Internet archives. <http://netpreserve.org/publications/iipc-r-003.pdf>
- [15] Internet Archive Wayback Machine. <http://web.archive.org/>.
- [16] Internet Archive. 2011. Wayback—introduction. <http://archive-access.sourceforge.net/projects/wayback/>.
- [17] Internet Archive. 2011. Wayback—requirements. http://archive-access.sourceforge.net/projects/wayback/administrator_manual.html.
- [18] ISO. 2006. ISO/DIS 28500: Information and documentation—the WARC file format. http://bibnum.bnf.fr/WARC/warc_ISO_DIS_28500.pdf.
- [19] ISO. 2011. ISO 28500:2009: Information and documentation—WARC file format. http://www.iso.org/iso/catalogue_detail?csnumber=44717.
- [20] Kahle, B. 2002. Preserving the Internet. *Scientific American* special online issue, 2 (Apr. 2002), 5-6. http://antonietta.philo.unibo.it/IUcorso2006-07/materiali/future_of_web.pdf.
- [21] Larivière, J. 2000. *Guidelines for Legal Deposit Legislation*. UNESCO, Paris. <http://unesdoc.unesco.org/images/0012/001214/121413eo.pdf>.
- [22] Living Web Archives (LiWA). <http://www.liwa-project.eu/>.
- [23] Longitudinal Analytics of Web Archive data (LAWA). <http://www.lawa-project.eu/>.
- [24] Masanès, J., Ed. 2006. *Web Archiving*. Springer-Verlag, Berlin Heidelberg.
- [25] Memento. <http://www.mementoweb.org/>.
- [26] Pop, R., Vasile, G. and Masanès, J. 2010. Archiving web video. In *International Web Archiving Workshop IAWA 2010*, Eds. J. Masanès, A. Rauber, M. Spaniol. <http://iwaw.europarchive.org/10/IWAW2010.pdf>.
- [27] Researchers and UK Web Archive Blog. <http://britishlibrary.typepad.co.uk/webarchive/>.

- [28] Society of American Archivists. 2011. So you want to be an archivist: an overview of the archives profession. <http://www2.archivists.org/profession>.
- [29] Thomas, A., Meyer, E.T., Dougherty, M., Van den Heuvel, C., Madsen, C. and Wyatt, S. 2010. *Researcher Engagement with Web Archives: Challenges and Opportunities for Investment*. JISC, London.
- [30] UK Web Archive. <http://www.webarchive.org.uk/ukwa/>.
- [31] UK Web Archive. 3D wall for special collections. <http://www.webarchive.org.uk/ukwa/wall/Blogs>.
- [32] UK Web Archive. 2010. Tag clouds. <http://www.webarchive.org.uk/ukwa/cloud>.
- [33] Wikipedia. 2011. List of web archiving initiatives. http://en.wikipedia.org/wiki/List_of_Web_Archiving_Initiatives.
- [34] Zotero. <http://www.zotero.org/>.
- [35] Zotero. 2007. Zotero and the Internet Archive join forces. <http://www.zotero.org/blog/zotero-and-the-internet-archive-join-forces/>.