JISC

Researcher Engagement with Web Archives
State of the Art

August 2010

Meghan Dougherty
Eric T. Meyer
Christine Madsen
Charles van den Heuvel
Arthur Thomas
Sally Wyatt

oii oxford internet institute university of oxford

Maastricht University

virtual knowledge studio

Acknowledgements

This report was funded by JISC, the Joint Information Systems Committee, from April to August 2010. The project was a partnership between the Oxford Internet Institute at the University of Oxford in the United Kingdom (http://www.oii.ox.ac.uk) and the Virtual Knowledge Studio at Maastricht University in the Netherlands (http://virtualknowledgestudio.nl/). Questions or queries about this report may be directed to:

Dr. Eric T. Meyer, Project Director
Oxford Internet Institute, University of Oxford
1 St Giles, Oxford, OX1 3JS, United Kingdom
Tel: +44 (0) 1865 287210
Email: eric.meyer@oii.ox.ac.uk

Neil Grindley, Programme Manager
JISC, Digital Preservation & Records Management
1st Floor Brettenham House (South), 5 Lancaster Place, London, WC2E 7EN, United Kingdom
Tel: +44 (0) 203 006 6059
Email: n.grindley@jisc.ac.uk

Please cite this report as:
Dougherty, M., Meyer, E.T., Madsen, C., van den Heuvel, C., Thomas, A., Wyatt, S. (2010). *Researcher Engagement with Web Archives: State of the Art.* London: JISC

Table of Contents

Boxed highlights in bold

Page is intentionally blank

Executive Summary

In this report, we summarize the state of the art of web archiving in relationship to researchers and research needs. This is a different focus than much of the earlier work in this area, including the JISC PoWR report which focused on institutional strategies for archiving web resources (JISC, 2008). It is important to note that this report focuses on the uses and needs of individual researchers. Research groups are also important, as some of the challenges that face individual researchers can quickly spiral into deeply complex tangles when dealing with collaboratories. For instance, national selection policies and national copyright rules can stand in the way of international projects, even if there are sound academic reasons to pursue international collaboration. While these issues are addressed here when appropriate, the bulk of the report focuses on individual researchers and institutions.

One of the main issues underlying this report is that there is still a gap between the *potential* community of researchers who have good reason to engage with creating, using, analysing and sharing web archives, and the *actual* (generally still small) community of researchers currently doing so. In this report, we identify some of the main reasons for archiving web pages, web sites, web domains, and the web in general. Beyond the fact that the web is allowing for the constant creation and distribution of huge volumes of information, it is also a valuable resource for understanding human behaviour and communication in the late 20th and early 21st centuries. To really reach the potential of web archives as objects of research, however, it is necessary to begin to take web archiving much more seriously as an important element of any research programme involving web resources.

A number of approaches are possible within this realm, and in the report we identify the differences in scope and scale of web archives, and present examples of how web archives can be used to address a number of research questions. Another key theme throughout are the challenges that still face researchers who wish to engage seriously with web archives as an object of research.

This report also makes a number of recommendations regarding developing additional capacity for web archiving and for research into web archives. These recommendations are grouped into three themes: building community, building tools & resources, and building practices.

Building Community

- Encourage the creation of communities that increase the accessibility and usability of web archiving tools
- Sharing tools and sharing web archives should become the norm
- New multidisciplinary approaches should be encouraged
- Privacy and property issues should be made more understandable
- Local instances of collections should feed into meta-collections to maximize the value of consortia

Building Tools & Resources

- There are two related and connected streams of support required to build infrastructure and to support the needs of individuals to archive

- Tools should be sharable and easy for researchers and librarians to implement
- Efforts should be made to diversify tool and interface development beyond preservation and into use
- Workflow tools should be used to orchestrate collections of standardised building blocks
- Tools should be developed that are able to execute query searches over multiple web archives
- Shared typologies, or vocabularies, of metadata need to be developed
- Standards, protocols and methods of quality control are need for interoperability, but not at the cost of flexibility
- Multiple access points into archives are needed to support administrative, descriptive, and conceptual access to web archives
- Shared archives of web archives need to be developed

Building Practices

- Web archiving needs to be integrated into the practices of institutions
- Additional training to understand the structure of web content will help researchers understand how to make use of archival web content in their research
- The possibilities of web archives should be communicated to a much broader research community
- Researchers need help to better match available tools to their needs
- Funding postgraduate students in areas that require web archives and providing them with the necessary skills will yield growth in this area in the long term
- Support for experimentation with web archives is vital for innovation
- Mentorship of new researchers is important for instilling the importance of archiving the web materials that researchers are increasing using as objects of study
- Measuring the impact of shared web archives is good practice

These recommendations are described more fully in the body of the report. We hope that these recommendations will be taken seriously, and that they will inspire researchers to see the advantages of working with web archives for research purposes.

Why archive the web?

The World Wide Web provides unprecedented access to information on virtually every known topic, and is a constantly growing and evolving information source that continues to develop as users and consumers of information and technology become increasingly knowledgeable. Information distributed on the web encompasses a vast array of the activities and artefacts of humanity. The *New York Times* reported in 2006 that the extent of human knowledge is summarized in "32 million books[1], 750 million articles and essays, 25 million songs, 500 million images, 500,000 movies, 3 million videos, TV shows and short films, and 100 billion public web pages" (Kelly, 2006). At the time, it was estimated that the sum of knowledge generated throughout human history could be contained in 50 petabytes (10^{15}) of storage space. The Internet, however, is increasing the rate at which textual, visual, and audio information is being produced and shared. By 2008, Google reported that their systems had found 1 trillion (10^{12}) unique URLs on the web at once (Alpert & Hajaj, 2008). The Internet Archive, which is a collection of historical copies of web pages representing the most complete source of the history of the Internet to date, currently contains 3 petabytes (10^{15}) of data, and is growing at a rate of 100 terabytes (10^{12}) of archived data *each month*. The sheer quantity of data appearing on the web represents a rapid expansion in human knowledge, including a comprehensive record of information production and social interaction over time. As Dr. Kirsten Foot put it when interviewed for this report:

> *At this point in our social material history, the extent of intertwining between online and offline phenomena is so thorough…that if we don't capture the online phenomena in at least the same rigor that we archive newspapers and other kinds of artefacts of cultural significance, we will have nothing to study retrospectively. There is a significant collective consciousness that is heading to a dark ages where we aren't writing anything down, in fact we are writing lots down on the web, but then we are writing over what we just wrote. It will be very hard for future scholars even in five years, ten years to understand what kinds of political and social and cultural moments or phenomena retrospectively without the key aspects of the web.* (Foot, personal communication)

What is a web archive?

In interviews for this research, stakeholders suggested the following answers to the question *"What is a web archive?"*

- A set of web objects that have been collected and verified with a particular purpose or goal in mind (where the goal could be to collect everything). What makes it an archive is the intentionality, collecting process, and then some level of verification
- Artefacts that are born digital, created on the web for the web, and are interesting for curatorial or analytical reasons
- A web archive is any offline storage of web content, created either manually or with an automation tool by an individual or group of people
- An accessible archive is one that has an interface that allows users to see objects in the archive
- A national collection representing website materials of interest to a nation
- A domain collection (e.g. ac.uk)
- A specialist collection based on one or more related specialist subjects
- A records management solution for business and legal purposes (one that treats a website as an organisational record)
- A collection designed to provide content of value to researchers (once one knows who the user community is)
- A collection of data that could be text-mined, or analysed statistically, or in other ways, to give interesting results
- A history of website design and application usage

[1] Although more recently, the Google Book project estimated the total number of books at a much higher count of approximately 130 million (Taycher, 2010). Estimates of this sort from any source are bound to be inaccurate in one way or another, if one wishes to take into account all languages at all times, but they can give one a sense of the scale at which one is operating when dealing with this much information.

This constant change is one of the web's greatest advantages to its end users: consumers of information are able to find the most up-to-date news and information at the touch of their fingertips. Yet this changing nature is also one of its chief frustrations as a data source: pages disappear, content is re-edited, comments are deleted, and wikis are vandalized. Without printed volumes, the history contained within the content of web pages is often lost. Researchers, archivists, librarians, students, citizens and corporations seeking knowledge or records previously but no longer available on the Internet are often at a loss, and those needing to know the history of content on the web are likely to struggle to get any significant information. Over the past fifteen years, most of the content of the web has disappeared as it is replaced by new pages and new content. There is, in fact, rapid turnover: several studies found that within a given week 35-40% of web pages changed their content[2] (Cho & Garcia-Molina, 2000; Fetterly, Manasse, Najork, & Wiener, 2004), and that this change is even more rapid when looking at subjects visiting dynamic pages such as news sites. For instance, in one study, 69% of web sites changed when revisited after a day or more (Weinreich, Obendorf, Herder, & Mayer, 2008), and another found that certain dynamic information is likely to change more frequently than once an hour (Adar, Teevan, Dumais, & Elsas, 2009). Pages are updated and refreshed continuously, but older versions are rarely archived by content producers. Web pages decay over time, and on average have a half-life of little more than two years, depending on the type of content (Koehler, 2004). This evolution and decay of content further results in a phenomenon referred to as 'link rot' as relationships and connections between data are lost over time (Taylor & Hudson, 2000).

In addition to this ever-changing content, the Internet and the web continue to show a dizzying pace of technological evolution – new multimedia types, new ways of displaying content (e.g. on mobile, rather than PC-based, platforms), increasing use of executable content such as JavaScript -- all pose new challenges for the web archive community. Worse still, much of the web's content (up to 90% by some estimates) is increasingly hidden behind forms-based query interfaces, and the actual content is held in databases which are inaccessible to crawlers; the development of methods to allow these "deep Web" contents to be collected poses another major challenge. Other, even more fundamental changes, such as the growing pervasiveness of social media such as Facebook and Twitter, among many others, point to a potential sharp decline in the relative importance of the "traditional" web, as is pointed out by in an article (Anderson & Wolff, 2010) which is engendering considerable controversy as this report goes to press. In this new world, there is a risk that open content, protocols and interface behaviours will be replaced by closed systems, content and interactions which are absolutely invisible to traditional archiving practices.

[2] Although the rate of change varied considerably by domain: .com pages changed much more quickly than .edu pages, for instance.

Web archives as research objects

Starting in the mid-1990s, researchers began partnering with librarians, as well as working on their own, to create archives of web objects that could be queried to draw generalizations about a variety of topics in the humanities, social, and physical sciences. Research using these methods range from studies about politics on the web (Foot & Schneider, 2006), to explorations of the web presence of different cultures (Franklin, 2005), to linguistic studies (McEnery & Wilson, 2001). These types of inquiry have contributed to shaping the descriptive, methodological, and theoretical bases of scholarship centred on web archives.

Web archives case: Election Web Spheres

Foot & Schneider's work (2006) was one of the earliest innovative research projects to use purpose built web archives as a means of answering a research question. In building their archive of web campaigning in the 2000, 2002, and 2004 elections in the United States, they conceptualized their objects of study as a *web sphere*. They define web sphere as *"a set of dynamically defined, digital resources spanning multiple web sites deemed relevant or related to a central event, concept, or theme…enabling analysis of communicative actions and relations between web producers and users developmentally over time"* (p. 27). By building an archived collection of websites produced by a variety of political actors during election campaigns, Foot & Schneider were able to better understand campaign strategies, tensions within campaigns, and more generally how technology is influencing the practice of political campaigning.

In the early 2000s, as web archives became more accessible and more widely known, a number of researchers and librarians worldwide began to investigate the potential and the limits of such a resource as a complement to exploration of the live, or currently active, web. Advocates of web archiving draw on methods in the relatively new area of digital cultural heritage to harness the quantity and variety of data available, in the hopes of advancing the potential for studying new genres such as blogs, web forums, and collections of emails. It is also possible using these methods to observe change in the content of the web as it takes place (Foot & Schneider, 2006; Kilgarriff & Grefenstette, 2003). Some sceptics, however, have questioned the trustworthiness of archives collected by researchers, arguing that control over sources and long-term replicability and stability in the building of such collections should be better defined (Brügger, 2005).

While many debates about the potential uses of web archives still remain at both a theoretical and practical level, web archiving is increasingly accepted by most cultural heritage institutions as an important complement to more traditional forms of collection development. Many researchers, too, have moved forward to explore the building and the resulting value of such archived web collections empirically. The development of social actions have been explored with the use of web archives (Foot & Schneider, 2006), object-oriented approaches in web historiography have been compared to topic and event oriented approaches (Dougherty, Schneider, & Jones, 2010, Forthcoming; Schneider & Foot, 2010), the ethical and legal impacts of saving artefacts from a highly volatile semi-public cultural space have been addressed (Dougherty, Foot, & Schneider, 2010). Within this body of work, technical and methodological approaches vary substantially: from the use of Google queries to derive artefacts from a web sphere to capture and archive (Schneider & Foot, 2004), and expert derived sets of artefacts to archive from the entirety of the web, to more targeted approaches delineating very specific sets of carefully defined web objects such as pages or sites (Brügger, 2005), and downloading quick-and-dirty specialized corpora for evaluating the language of the web (see, e.g., the papers in Baroni & Bernardini, 2006). While this work has provided interesting tools and new insights, none so far has succeeded in coalescing and making available to the larger research and heritage community an infrastructure that combines the advantages of the web in terms of inclusion and access with the advantages of traditional methods in archive research in terms of stability and control.

This report presents an overview of the current state of web archiving, including the diversity of practices as they are evident in a variety of inquiry modes, attempts at standardization, and the

loose web archiving infrastructure that has emerged to support e-research and e-heritage. The focus of this project, though, is on the current state of researcher engagement with web archives – how are researchers currently making use of web archives and what sort of technical and policy infrastructures will they need in the future in order to facilitate their work?

State of the art

Stewardship of cultural heritage is a story of loss and reconstruction. Artefacts deteriorate, or become otherwise corrupted, and stewards of the cultural heritage those artefacts represent - whether they be scholars, curators, archivists, or interested amateurs - feel a responsibility to reconstruct not only the artefacts, but often the meaning the artefact holds for interpreting our past. This holds true for stewardship of digital cultural heritage as well, not only in the construction of narratives about our past on the web, but also for the way practices are developed for handling the web artefacts that help researchers to construct those narratives.

The World Wide Web is now largely recognized as an essential access point for cultural, historical, and scientific information. Nonetheless, it is still a highly fragmented environment that is often changing, always evolving, and often disappearing. In recognition of this problem, several groups are now successfully archiving large portions or selected segments of the web. Through these activities, they aim to create an archival record of web culture or of contemporary culture as manifest on the web. This record is intended "to resemble a digital library" from which historians, curators and scholars can draw data to support their research (Lyman & Varian, 2003).

Library and information science have been developing practices for collection and archive development for decades that have come to dominate web archiving. In some ways, the practices and standards of this discipline are a good fit because they are extensively developed and ready to handle the content management and delivery systems required by web archives. Further, they offer an existing policy framework for the collection of contemporary cultural materials. However, there are consequences to relying heavily on libraries and archives to deal with web archives. As European Archive director Julien Masanès points out:

> *It is a utopia to hope that a small number of librarians will replace the publisher's filter at the scale of the global Web. Even if they have a long tradition in selecting content, they have done this in a much more structured environment that was also several orders of magnitude smaller in size. Although this is still possible and useful for well-defined communities and limited goals…, applying this as a global mechanism for Web archiving is not realistic. 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.* (Masanès, 2006, p. 4)

While library and information science norms have been the basis for many of the developments in web archiving policy and infrastructure, the resulting focus on collection development and preservation of artefacts has often been done with little regard to the question of how the web archives will eventually be used. Viewing the web archive as a collection of documents and bibliographic records is an efficient approach to storing and preserving the web. Whether it is flexible enough to accommodate the uses that researchers will want to put web archives to is another question. This has set up a point of contention between librarians and information scientists who would like to build widely valuable and accessible collections, and humanities and social science researchers who would like to develop web archiving as a method for understanding digital cultural heritage or web historiography. The two perspectives are not diametrically opposed, but there are certainly points of contention that are derived from differently held philosophical undercurrents that motivate each (Dougherty, 2007). Librarians and archivists are inclined (and trained) to build

Web archives case: Iranian Elections

In June 2009, Iran participated in its tenth democratic presidential election. As the results were tallied, allegations of electoral fraud were voiced and protests mounted. Most of the anti-Ahmadinejad actions known as the Green Movement were coordinated online. According to one researcher, "*Immediately after the election there were lots of digital materials online – campaign materials, online activism, video clips, citizen journalism, and a lot of really good stuff in Facebook. Essentially there was a huge amount of Iranian cultural artefacts online. Nothing like this had ever happened before.*"

A group of researchers distributed around the world attempted to archive these materials. They had two motivations: *"The first is selfish, really. That these would make a great research archive at some point. Something to go back to. The second is political. Through this archive it would be easier to reproduce the narrative of the Green Movement."*

Unfortunately, the project ran into technical problems due to a lack of easy to use tools and server space. Without an immediate source of funding to pay for commercial services, the researchers were not able to save most of this material. This underscores the need to have better and more accessible methods to archive and save materials related to unfolding events that are now being lost.

collections that will last for centuries, even 'forever' as is the mandate for some institutions. Researchers are interested in first building or collecting something that can help them answer their current research questions or even design new ones. The longevity of the data beyond their own career or even beyond a project, for researches, is generally of secondary importance.

Collaboration and partnership is a complex issue. During an interview for this project, Kirsten Foot reflected on the issues that arise in institutional collaborations. She identified the various partners who are interested in partnering around web archiving: national libraries in the US, Europe, Australia and Asia; and museums and archives that are recognizing the value of born digital objects for their collections. She mentioned universities as institutions that are taking an interest, but quickly explained that they do not seem to have yet developed any discernible strategy for collecting born digital materials. In describing her experience as an academic entering into a multi-institution partnership, she explained that even as an individual researcher, there needs to be some university-level commitment to support inter-institutional web archiving activity. She mentions that there are legal considerations as well as technical considerations that serve as a foundation. The more complicated issues are the detailed protocols about what curation consists of, and the basis for collection development. These issues are approached very differently by social researchers versus librarians versus archivists. Foot said, "It is important to really thrash through those [differences] and work out a protocol." Technical questions of storage, quality assurance, and capture are also issues to be negotiated. When Foot was asked to elaborate on "thrashing through differences" to determine protocol, she said that she learned the hard way that these are necessary conversations. People from different types of disciplines have different concepts in mind even when using the same terms and it is important to bring those differences to the forefront when collaborating. She was particular about the definition of what it means to be "systematic" and the level of rationale or criteria needed to complete a given project successfully. There are different practices from professional communities and domain expertise. Thoughtful agreement around these issues are increasingly important as proprietary technology can obscure how we access and capture web materials - different search engine algorithms will lead to different results much the same as different search strategies will surface different results. These differences have deep epistemological and disciplinary roots.

As a result, large libraries and archives continue with their efforts to build large multi-purpose web archives that further institutional missions, while researchers - either on their own, or partnering with archivists - develop their own archives for use in their research. Archives cannot justify allocating resources to project-specific archives, but researchers cannot always find useful materials for their work in the large multi-purpose archives being built by archivists. The core tools for creating basic web archives are now widely in use, but there is no underlying infrastructure in place to support the research into these archives.

Consequently, web archiving is currently in a state of flux where boundaries around traditional roles of researchers and stewards are blurring. Stewards are seeking out researchers to learn their needs. Researchers are building their own collections and seeking the expertise of archivists to sustain those collections. These types of collaboration are resulting in the need to experiment with different approaches that are guided by multiple motivating principles. Web archives created by a social scientist will inevitably differ from those created by a librarian, or by a linguist. The tools needed to make the archives usable to each group will vary as well. Each practitioner is motivated by a different mission, be it institutional, methodological, or epistemological. Diverse approaches to web archiving are resulting from this experimentation and are increasingly leading to conversation and collaboration across fields to develop inclusive practices.

In addition to this older community, who have been principally interested in the content of the web, we now see the appearance of a relatively new community – the Web Scientists – who are interested in the web itself as a technological artefact and object of study (Hendler, Shadbolt, Berners-Lee, & Weitzner, 2008). There are many fascinating issues about the network structure of the web, and the ways in which that structure evolves over time, which have intrinsic interest, as well as telling us a good deal about how human beings use communication technologies in innovative ways to interact and collaborate in the creation of new cultural artefacts. The interests of these new students of the web are not necessarily best served by the library and information science approach, and future developments in web archiving will need to take these new requirements into account.

A diversity of approaches distinguished by purpose

Each of these diverse approaches to archiving web objects develops from certain modes or styles of inquiry. Researchers in the social sciences and humanities are guided in their practices by methodological concerns and specific research questions when approaching the web and attempting to stabilize objects of analysis there. Cultural heritage professionals are guided by institutional mission statements and clientele.

The greatest contention among these professionals is based on fundamental differences in how we understand the world, and how we determine what things are. These epistemological and ontological beliefs provide a driving force for activities of collection, documentation, classification, and are eventually filtered through to defining points of access. Divergences in the beliefs that underscore the development of these activities can entrench practices later, so much so that change becomes quite difficult. Support for experimentation in practices is vital at these early stages as the field is still being defined.

The following categorization of web archiving projects is not comprehensive, but shows the evolution of multiplying practices and tools. Each step problematizes the previous one and creates its own new path while respecting the value of the previous. Each new path proposes its own set of practices as an addition to add value to previous collection practices.

Broad collections: diverse future uses

Both scholars and cultural heritage institutions recognize the need and value of preserving content on the web (e.g., Arms, Adkins, Ammen, & Hayes, 2001; Burner, 1997; Day, 2003; Foot & Schneider, 2002, 2006; Hodge, 2000; Kahle, 1997; Kahle, Prelinger, & Jackson, 2001; Lyman & Kahle, 1998; Masanès, 2002, 2005, 2006; Schneider & Foot, 2002, 2004, 2005), and have launched efforts to archive web content.

In 1997, Brewster Kahle published a short article in *Scientific American* entitled, “Preserving the Internet” in which he described his Internet Archive project that would attempt to do just that. This

was not the first or only mention of web archiving at the time – the Finnish EVA project was launched in the same year and Australia's PANDORA archive was launched in 1996 – but it marked the beginning of the most ambitious effort to preserve artefacts from the web to date. The Internet Archive (IA)[3] takes a whole-domain approach, with the goal to preserve the entire content of the global web. This approach builds a comprehensive collection of websites and online resources using harvesters to automatically retrieve artefacts in broad sweeps of the web. Other broad sweepers of the web include the European Archive[4], while projects such as the Swedish Kulturarw3[5] and the UK Web Archive[6] limit their domain to national web spaces. The Preserving Access to Digital Information (PADI) page[7] maintains a list of national web archiving programmes.

Broad scale collecting strategies result in very large collections of archived sites, but generally with little documentation or metadata about the objects. Due to the sheer scale of IA's crawls, for example, only machine readable data is collected. This results in archives that are difficult to navigate as archived sites can only be retrieved via URL, as in the IA's case via their *Wayback Machine* interface[8]. This interface problem is exacerbated by the fact that the quality and reliability of these archives often do not meet the standards of completeness and replicability required of researchers in the humanities and social sciences. However, new tools from IA such as *Archive-It*[9] are being developed to allow for more focused collections with advanced features such as search, a feature which is not yet technically feasible across the entire *Wayback* collection (see box). As will be discussed further below, access, interfaces, and selection policies are all creating challenges for those wishing to broaden the use and re-use of web archives.

Web archives challenge: Search

In 2009, the Internet Archive ran a pilot in providing full-text searchability, making the first five years of their archive (1996-2000) available for searching. The search ranking mechanisms available at that time were not adequate, however, and the search results were full of spam. To date, there is still no reliable full text search tool for web archives and although several groups are currently working on the problem, it remains one of the greatest obstacles to providing archives usable for a wide variety of researchers.

Search in general is still not able to adequately work with items in digital archives to the standard many researchers desire. For instance, with regard to the *New York Times* digital archive of news content dating back to 1851: *"We can say, 'show me all the articles about Barack Obama,' but we don't have a database that can tell us when he was born, or how many books he wrote… Such a resource will not only help the research community move the needle for our company but for any company with a large-scale data-management problem."* (Evan Sandhaus, New York Times Research and Development Labs, quoted in Simonite, 2010)

In addition to building collections, large-scale projects such as the Internet Archive and the European Archive have parallel missions to make their collections usable and accessible to the public and to researchers. For the former this is focused on universal accessibility--that is, to the widest audience possible. To date, their efforts have been primarily focused on providing "native replay" of individual archived sites and pages. With this capability now well established they are turning their attentions to providing new ways for researchers to use their archive (primarily through the development of new APIs). The European Archive, too, is focused on building tools that allow researchers to engage with their archives, for example to run analytics or to perform linguistic analysis. Through their *Living Knowledge* project[10] their goal is "goal is to bring a

[3] http://www.archive.org/
[4] http://www.europarchive.org/
[5] http://www.kb.se/english/find/internet/websites/
[6] http://www.webarchive.org.uk/ukwa/
[7] http://www.nla.gov.au/padi/topics/92.html
[8] http://www.archive.org/web/web.php
[9] http://www.archive-it.org/
[10] http://livingknowledge.europarchive.org/index.php

new quality into search and knowledge management technology for more concise, complete and contextualised search results."

A number of studies have established the Internet Archive as a valid tool for research in the social sciences. In particular, scholars have used the Internet Archive's *Wayback Machine* as a tool for estimating the age of a website, the frequency of updates, and for evaluating and coding the content within sites (Brock, 2005; Thelwall & Vaughan, 2004; Veronin, 2002). Further, Murphy, Hashim & O'Connor (2008) validated measures of age and frequency of updating against third-party data, illustrating the overall strength and reliability of these measures as research tools. Thus, there is support for the use of data from the Internet Archive as attributes and characteristics in research studies. The *Wayback Machine* can additionally be used as an evolutionary research tool to track the development of technology over time, for instance, to track changes in content over time. Chu, Leung, Van Hui & Cheung (2007) conducted a longitudinal study of e-commerce websites, using the Wayback machine to track the development of site content. Similarly, Hackett & Parmanto (2005) used the *Wayback Machine* to analyze changes in website design in response to technological advances over time. Efforts along these lines include the *Memento* project[11], which adds a time dimension to the HTTP protocol to better integrate the current and past web, and the *Yahoo Time Explorer*[12] which is being developed to build timelines from searches in news archives. A number of scholars have conducted historical research using data from the Internet Archive. This previous work has clearly established the utility of data from the Internet Archive as a source of research data. Yet large-scale studies using this source are hampered by the size of the database, the structure of the data itself and the complexity of linkages between sites (Murphy, et al., 2008). To date, they have used tools that have been time-intensive to develop, that are custom-made for particular topics and therefore not widely usable, and that have encountered many other difficulties and limitations.

Web archives case: The Twitter archive

In 2010, the U.S. Library of Congress announced that Twitter had given its entire archive of public tweets to the Library for preservation and to make it available for research use. According to the FAQ for the collection, *"Twitter is part of the historical record of communication, news reporting, and social trends – all of which complement the Library's existing cultural heritage collections. It is a direct record of important events such as the 2008 U.S. presidential election or the "Green Revolution" in Iran. It also serves as a news feed with minute-by-minute headlines from major news sources such as Reuters, The Wall Street Journal and The New York Times. At the same time, it is a platform for citizen journalism with many significant events being first reported by eyewitnesses. The Library of Congress collections include items such as the very first telegram ever sent, by telegraph inventor Samuel F.B. Morse, oral histories from veterans and ordinary citizens, and many other firsthand accounts of history. These collections and others have left behind glimpses of the lives of ordinary people, thereby enriching knowledge of the context of public events recorded in government documents and newspapers. Individually tweets might seem insignificant, but viewed in the aggregate, they can be a resource for future generations to understand life in the 21st century."* (Raymond, 2010)

Directed collections: flexible, immediate uses by individuals and institutions

Other web archiving approaches are selective, thematic, deposit-based or a combination of these approaches. Selective approaches identify web artefacts to collect by specifying certain inclusion criteria such as a theme, by quality or significance, or through identifying specific intervals at which to take impressions or snapshots of web artefacts. This type of selection at the harvesting level is employed by Australia's PANDORA[13] project, which collects selected Australian online publications deemed to be of national significance and long-term research value. The U.S. Library of Congress employs a thematic approach with its Library of Congress Web Archives[14] (originally called the

[11] http://www.mementoweb.org/
[12] http://fbmya01.barcelonamedia.org:8080/future/
[13] http://pandora.nla.gov.au/
[14] http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html

Web archives methods: Collecting

Ed Pinsent of the University of London Computer Centre provided the following general steps he uses in creating a web archive.

1. Discover that the target site exists - for example by checking jisc.ac.uk and other sources to see what new projects have started up, whether they have websites, and determine if they fit the scope of the collection.
2. Seek permission from the website owner to make a copy. I use a form and mail merge to do this. The project manager is usually regarded as the owner. If and when consent is given, enter the details of their Institution into Web Curator Tool, thus creating a permissions record.
3. Create a target entity in Web Curator Tool and link it to the permissions record.
4. Set harvest in motion.
5. QA the results. If necessary, change parameters of the harvest for future gathers (e.g. add or remove filters), or "prune" the gather to remove material we don't need
6. Submit the harvest to the archive.

MINERVA project) by selecting artefacts that fit a specific theme. Its *United States Election 2000 Web Archive*[15] (also done in 2002[16], 2004[17] and 2006[18]), *September 11, 2001 Web Archive*[19], *Iraq War, 2003 Web Archive*[20], and *Papal Transition 2005 Web Archive*[21], for instance, used these themes to guide selection. Deposit-based projects, such as projects at the National Library of the Netherlands (Koninklijke Bibliotheek)[22], rely on voluntary deposits of web artefacts. The National Library of the Netherlands is also working with experts on collection strategies within specific identified humanities-related topic areas.

Several projects aimed at preserving national digital cultural heritage employ a combination of these approaches. France and Denmark combine comprehensive sweeps with targeted selective and thematic collection strategies in an effort to guarantee good coverage of certain highly valuable portions of web artefacts within a larger broader sweep of content. The Digital Archives for Chinese Studies[23] (DACHS) with branches at the University of Heidelberg and Leiden University, and Virtual Remote Control[24] (VRC) at Cornell University represent a 'by discipline' approach to web archiving that is popular among research institutes and universities. The British Library takes a similar hybrid approach, focusing on building discrete collections of "websites with research value that are representative of British social history and cultural heritage".[25] Several of Harvard University's libraries[26] are working on very narrow but deep collections, known to fall within the existing collection scope of the library, such as *Blogs: Capturing Women's Voices*[27] and the *Constitutional Revision in Japan Research Project*[28]. At both the British Library and Harvard University Library, archiving of web content is being integrated with standard collection development practices. These approaches provide varying degrees of nuance in all the processes of web archiving. Libraries, archives and large cultural heritage institutes can have broader objectives and thus employ broader practices in their approaches.

[15] http://lcweb2.loc.gov/diglib/lcwa/html/elec2000/elec2000-overview.html
[16] http://lcweb2.loc.gov/diglib/lcwa/html/elec2002/elec2002-overview.html
[17] http://lcweb2.loc.gov/diglib/lcwa/html/elec2004/elec2004-overview.html
[18] http://lcweb2.loc.gov/diglib/lcwa/html/elec2006/elec2006-overview.html
[19] http://lcweb2.loc.gov/diglib/lcwa/html/sept11/sept11-overview.html
[20] http://lcweb2.loc.gov/diglib/lcwa/html/iraq/iraq-overview.html
[21] http://lcweb2.loc.gov/diglib/lcwa/html/papal/papal-overview.html
[22] http://www.kb.nl/hrd/dd/dd_projecten/webarchivering/index-en.html
[23] http://www.sino.uni-heidelberg.de/dachs/
[24] http://handle.library.cornell.edu/VRC/
[25] http://www.bl.uk/aboutus/stratpolprog/digi/webarch/index.html
[26] http://wax.lib.harvard.edu/collections/home.do
[27] http://wax.lib.harvard.edu/collections/collection.do?coll=61&lang=eng
[28] http://wax.lib.harvard.edu/collections/collection.do?coll=101&lang=eng

Narrow collections: known, immediate uses by researchers

Oftentimes, a researcher's systematic approach and sometimes-narrow topical scope guides the creation of narrow collections in web archiving. In these researcher-led cases, the selection of artefacts is driven by the boundaries of the research project for which a sampling scheme has been developed. Categorization follows coding strategies informed by prior inquiry into the field and developed to address certain concepts tested in the project. These collections are limited in size and scope. They typically focus on an initial list of seed URLs, or the contents of one website, and contain frequent (sometimes hourly or more) captures of artefacts resulting in very full, but limited collections. There are sound methodological reasons for creating a web archive; as project interviewee Dr. Steve Schneider, of SUNY Institute of Technology in the United States, puts it:

> *I think it is not possible to study social phenomenon on the web, especially in an ongoing/ developmental sense, at any medium-to-large scale and with any hope of replicability, without archiving material. So the benefit is that archives make it possible to do the quality social science research that is, in a sense, competitive methodologically with large-scale survey research. My thoughts are that the way we approach web sphere analysis has the opportunity to bring the methodological sophistication (including the ability for others to replicate our research) of public opinion research to the study of online social phenomena.* (Schneider, personal communication)

Some of the more technical aspects of web archiving such as indexing and curation are similar in both widely sweeping archiving schemes and narrowly bound scholarly web archiving. Scholarly web archiving is a focused development of a collection following narrowly defined collection strategies, while individually produced web archives are designed to be a source of data generally for one particular project. Researchers develop these collections on their own, and in conjunction with larger institutions with better resources, however the extent to which these collections can be described as archives varies. Individual collections with no public access and no claims to longevity can hardly be called archives, but this does not reduce their potential value to the research community. They merely lack infrastructural support.

Traditional collection development can follow similar individual procedures, but without a specific research project in mind. Collection development is an ongoing task that follows set policies, but is a different act than the sampling procedures in a research project that tend to guide scholarly web archiving collection development. Fundamentally, the archivist aims to develop a collection that may be widely used for any number of known and

Web archives case: Immigration web storm

Interviewee Dr. Kirsten Foot of the University of Washington in the USA, recently compiled a web archive of what she calls a web storm, which she defines as *"a flurry of productive activity that happens on the web in unpredictably predictable ways. You don't know when it will happen, but there will be bursts of generative activity on the web in which many actors are producing material about the phenomenon."* As a social science researcher studying social phenomena, she often has an eye out for unanticipated web storms that fit into other arguments that she is interested in theoretically.

This particular case involved the Yahoo News site and the recent immigration debate taking place through links to Photoshopped images of a particular cartoon character. Foot noticed that Yahoo News was aggregating reports from other news sources reporting the photo manipulation as political commentary, but were presenting the content on their site in a guarded way. She noticed that Yahoo News was providing access to the politically and emotionally charged images through a link to an outside server and providing their own disclaimer in text surrounding the link on their page. Foot saw this as an example of strategic coproduction, and began capturing snapshots of the Yahoo pages, and its target links. Once she noticed what was happening and identified the event as an example of a concept she works with, she explains that she knew there were certain aspects of the phenomenon that she needed to capture on the pages that were linked together in this event. She needed to capture evidence of the particular dimensions she saw as relevant to the concept she was observing: who was hosting the images, who was pointing to them, what the various pages the portal provided were, the various levels that it took to navigate to the page with the image, etc.

unknown users for different purposes. The individual researcher's web archive is a set of data collected to support a specific inquiry that may be re-purposed for another project later.

One particular form of narrow archive is the “idiosyncratic archive" (of the type described in the accompanying box on the Immigration web storm, and also discussed in Dougherty, et al., 2010, Forthcoming). The web is often the site of "unpredictably predictable" activity, a type of activity that is not necessarily tied to the definition of a web storm presented in the box, but is an undercurrent concept that drives all activity and retrospective analysis on the web. It is degrees of this "unpredictable predictability" that illustrates the difference between different styles of narrow archives. So, for instance, there is a difference between the unanticipated web storm such as the Yahoo News example, and an event such as the recent case of Steven Slater, the JetBlue employee who dramatically quit his job by exiting the plane by the emergency chute. Though the time scale is still short, there is a moment - no matter how short - between the event of Slater's dramatic exit from his flight attendant job and the coming web storm for which you can predict what online actors will produce a short-lived burst of related content. In contract, Foot's Yahoo News web storm brews more slowly from events originating on the web.

Individual or research-led web archiving usually includes rich metadata, interpretation, and representation. These are technical and analytical steps that actively engage the user or reader. These steps go beyond other methods of web archiving by invoking research methodology designed to answer specific questions, rather than simply to catalogue and preserve information. This added data makes the resulting web archive particularly useful to the researcher or archivist who created it. The risk, of course, is that without an ontological understanding of those methods and collection development policies, these collections may be difficult for other researchers to use.

Page is intentionally blank

Web archiving: A developing field

No matter what the approach, web archiving is a complicated process involving many steps to selecting and acquiring objects to archive, and also determining solutions for storage, documentation and access (Arms, et al., 2001; Foot & Schneider, 2006; Hodge, 2000; Masanès, 2002, 2005, 2006). While there are many planning strategies and policy formats, collection development policy making takes knowledge, experience, and intuition, but it also aims to reflect the needs and interests of the collection's community of users (Johnson, 2009). Like most others defining the scope of web archiving, Julien Masanès, director of the European Archive, does so by developing practices already established in librarianship and archiving. The practices he describes cover collection policy development and collection building, but fall short of delving deeply into the other areas of access, categorization, interpretation, and representation.

Masanès (2002, 2006) points out that applying traditional strategies for collection management to web collections is difficult, a point also noted in interviews done for this report. At this point in time, there are few comparable collections against which to evaluate the completeness of web collection strategies. The inconsistent publishing procedures and formats on the web, and the connectedness of the medium both create a need for a different and more open approach to discovery and dynamic selection.

A cultural environment exists with this technical media environment that is also fluid. It is this type of context and environment that allow us to recognize artefacts and their uses. We use these environments to create genres into which we can categorize artefacts to account for their meaning and usefulness (Innis, 1951; Levinson, 1997). The preservation of a digital document is tied to its production. Every time you read a digital artefact, it must be reproduced and reconstructed entirely – it must be rendered in a human-readable format. With born-digital documents, preservation is no longer an artefact-centric problem. The integrity of the media environment surrounding and supporting the artefact must be preserved in addition to the integrity of the artefact.

Some web archivists discuss preservation, but their discussion of what they call preservation also addresses issues of selection and capture (Day, 2003). The rate of resource decay on the Internet, the rate of change in web tools and standards, and the continuing development of the Semantic Web, where information is given well-defined meaning so machines can recognize, understand and process it accordingly, are all issues to address when developing a collection policy, and they will influence choices of how to collect, when to collect and what to collect. None of these considerations address how to preserve the artefacts once they are collected, nor do they address how to preserve the varied uses and interpretations the artefacts took on during their active time in the cultural world (that active time may overlap with the time spent in the archive once collected). As one of the librarians interviewed for this report said, “innovation in web technologies is both a challenge and a threat. We are always catching up.” The social life of the artefact, defined by the uses to which it was put to produce new knowledge and the interpretation it was assigned by different users at different times are additional avenues in which to collect metadata to preserve not only the object itself, but some meaning about the object so its cultural value can be revisited and evaluated as it changes over time.

Tools for building and using web archives

Each individual tool for personal desktop archiving has a different set of goals and so different design elements. Simply archiving sites you’ve visited during a particular research setting does not always meet the needs of the researcher. Often, the researcher does not know what metadata elements are missing, or what indexing elements are missing from a certain archiving tool until it is too late. Social science researchers find themselves with archives that are full of redundancies that need to

be cleaned out, missing seeming redundancies that actually show significant change, or contain a mess of archived sites with no logic of how the individual objects can be related to one another. Personal desktop archiving tools are designed from a "basic needs" perspective. The designer's assumption is that the user wants to save websites to view later. The next level of design complication is that the user may want to know the exact click stream followed when navigating a site. Neither of these assumptions touches upon the complexity of what a social science researcher thinks it means to save a website for retrospective study, or to archive a click stream for analysis. In an interview for this project discussed above, Kirsten Foot described problems in inter-institutional collaborations in web archiving; she explained that people from different disciplines have different concepts in mind even when we use the same terms. These differences can surface in the design of personal desktop archiving tools. It is important to surface those differences early. It is important for researchers to be very clear about their research goals, and thorough about what metrics they will need to reach their goals. It is also important to develop some tools that are not multi-purpose. Not all tools need to be accessible to the casual user, and special research tools can be designed to meet the higher level needs of the researcher.

Web archives case: Personal Facebook archives

Interviewee Frank McCown recently led a research project that produced a Facebook archiving add-on for Firefox (ArchiveFacebook, available at https://addons.mozilla.org/en-US/firefox/addon/13993/). The add-on is a tool that users can install and run by themselves to produce an offline, fully navigable archive of their Facebook account. This kind of individual-use tool reinforces the current trend of creating fail-safes and living-wills for online identity profiles. This is a specific perspective on archiving the web, which has potential to find a large popular following of users for this type of tool, but does not necessarily help researchers create, access, or analyze web material retrospectively. More often than not, this is the type of archiving tool that is leading the current state of web archiving tool development.

The overarching challenge is not recognizing the importance of archiving web content in general, or more specifically a particular metric, concept, or method until it is far too late. Certain questions cannot be answered, certain concepts cannot be illustrated, certain methods cannot be used if measures are not set up to be indexed as an archive is built, and studies cannot be replicated if the ephemeral digital primary materials aren't archived. Even if the researcher was clever or lucky enough to capture all the different data required, there are two additional challenges. The first is finding software that suits the researcher's needs, and as a corollary finding a researcher who is capable of evaluating the available tools to match their needs. It is hard to find and figure out which archiving software is going to be useful and user friendly for the kind of use in practice that that individual has. The second challenge in use is organization. The structure of the objects you collect matters. Foot described seeing eager researcher-archivists collect strategically, only to find that their collection was inaccessible due to tremendous redundancies, and structural chaos in the archive: "Many of the tools available are simply not robust enough" (Foot, personal communication).

In 2003, twelve institutions including the Internet Archive and eleven national libraries (Australia, Canada, Denmark, Finland, France, Iceland, Italy, Norway, Sweden, British Library, US Library of Congress) formed an international collaboration focused on Internet Archiving. The International Internet Preservation Consortium[29] (IIPC) focuses on creating tools and standards for web archiving as well as providing support and advocacy for its members. The IIPC open source tools now comprise the standard package used by most cultural institutions engaged in web archiving. These include the Heritrix crawler, the Web Curator Tool (WCT) for collecting, NutchWAX for indexing, and the Wayback interface for access.[30]

[29] http://www.netpreserve.org
[30] http://www.netpreserve.org/software/downloads.php

As described by one of the web archiving project managers from a national library, *"we have just now gotten good at what we do – downloading copies of static text from the web."* Nearly twenty years after the introduction of the first web browser, we have finally made progress in capturing and preserving some of the earliest web documents. This pace, when contrasted with the speed of innovation on the web, will shortly become a significant challenge facing web archiving communities. New protocols such as iPhone apps[31] are being introduced and popularized across the web. New mobile devices are providing new ways of looking at the same data, increasing the difficulty of providing "native replay" of archived materials. This begs the question: if some web based content appears differently in a web browser and on an iPhone, does a web archive need to capture that? If so, where does it stop – do archives need to capture all competing versions as well? Even the introduction of embedded metadata in to existing protocols (RDF, for example), which could help with indexing and access to pages within archives, provide new challenges. For example, if the public content of a web page does not change, but its tags do, does this represent a new version? As Wendy Gogel of the Harvard University Library commented to us for this project, *"In the future the sheer number of formats is going to be overwhelming and the problem is not the capture of these, but in being able to preserve display of them."*

Future challenges and opportunities for using web archives

Differing inquiry modes for web archives

As outlined above, there are several current approaches to building web archives - some arising from frustration with existing resources, some developing from institutional mission statements, and all developing from limited understanding of the end-users' needs.

Temporary *ad-hoc* practices that are developed to circumvent obstacles were discussed in several pilot interviews conducted in autumn 2008 by project partner VKS (Dougherty, 2008), and in new interviews for this project conducted in summer 2010 with a range of researchers and librarians engaged in some variation of web archiving. All respondents are facing similar sets of obstacles despite their approach. The ways in which these obstacles are handled determines, among many things, the character of the resulting archive, the limitations of use as set by access points to the resulting archive, and ultimately the perceived value the resulting web archive offers to different communities of researchers.

Common obstacles

The common thread through conversations among researchers and archivists using and building web archives is that researcher-users all want different aspects of the same things. Firstly, they want ***stabilized web objects*** that can be reliably studied and cited. They want to be able to ***clearly define*** what that stabilized archived object represents in reference to the live web. They want to have access to archived ***representations of the most fine-grained features*** of web objects in order to suit their research needs. Most of all, they want to ***work with those objects, enriching and annotating them*** on whatever level is appropriate for their analysis.

In terms of the archive itself, three things are clear: an archive must be trustworthy, long-lasting, and reliable. These are fundamental elements of any archive; and these elements need to be extended to bolster web archiving processes as they develop. Researchers and small-scale libraries are increasingly seeking the help of large established archives to meet these standards. Resources for downloading, archiving, and serving archived objects are often too costly to implement for individual researchers and small libraries. Even with the availability of software tools such as those provided by

[31] Short for applications. Generally small, inexpensive, single purpose programs designed to pull data for instant display.

IIPC[32], the limited access to human and technical resources and expertise is often cited as the main obstacle for small libraries and researchers wanting to participate more actively in the web archiving community. Even with free technical resources available, small operations have limited human resources to run and maintain it. These parties recognize that the criteria for legitimately calling their collection a valuable archive that serves a research purpose in the future are often beyond their reach. They are seeking to collaborate with larger archive institutions to share resources and expertise.

As small collections seek collaborative opportunities, they move forward, doing their best to meet standards of a legitimate archive, and face the next set of obstacles: **access**. Often, access obstacles are also fundamentally a problem with lack of resources. In the case of access, not only is there a lack of labour resources, there is also a lack of technical infrastructure to support that work.

For these archives to find value in the world or research, they need to have multiple access points: **administrative**, **descriptive**, and **contextual**. These types of access points are experimented with and employed in myriad ways in different archives. Again, there are few shared practices, and no standards across archives. Shared practices exist only as a coincidence if two archives use the same harvesting software, or object-rendering software. Further, these three access points are even described differently using disciplinary language that is not shared between researchers and archivists, or even between researchers in different fields. According to an archiving engineer we interviewed at one of the national archives, quality assurance still requires extensive manual work as few automated tools exist, exacerbating the problem since manual steps are more difficult to duplicate unless they are meticulously documented. Each description of how an archivist or researcher would like to have access to an archive contains elements of these three strata, but none share a common language.

Administrative access enables a user (or archivist) to examine an artefact and determine exactly what it is (when it was archived, with what software, from what organization, including what file types, etc). This type of access is imperative for the structure of the archive itself. Administrative data enables an archivist to rebuild an archive after a data crash, for example. Administrative access is also valuable for content comparison across archives or across archived objects.

Descriptive access is basic catalogue access to artefacts in an archive. Basic cataloguing information makes artefacts findable (Morville, 2005). This descriptive metadata is equivalent to the information in a library that would help a user find one book among many on a shelf. The metadata answers the question, “What is it?” for every object in the archive.

Contextual access places artefacts in a thickly described and purposeful context. Contextual access does not place an artefact in its original context; rather it makes an artefact findable via its relationship to other objects in a research project. Contextual access has been experimented with in several collections; two of the most notable are DACHS[33] and the former Politicalweb.info, which somewhat ironically is no longer available online.[34] Users enter an archive and view archived artefacts via the research of another. Archived artefacts, in this sense, can be seen as a collection of objects to which a research project refers. This metadata answers the question, “What is it *about*?” for any object in the archive, and this question can be answered differently many times over depending on the perspective and purpose of the researcher-user.

[32] http://www.netpreserve.org/software/downloads.php
[33] http://leiden.dachs-archive.org/
[34] See the Wayback Machine to view archived versions of the site: http://web.archive.org/web/*/Politicalweb.info. The domain name currently points to an advertising site.

A related issue is contextualization in the form of **annotation**. Hanzo Archive, for instance, has created tools that allow individual annotations to artefacts within a web archive. The need for collective annotation of web archives, however, is only recently being acknowledged (Dougherty, 2007; van den Heuvel, 2009). By allowing collective annotation of web archive objects, researchers can build up additional levels of data to enhance our collective memory (van den Heuvel, 2009, pp. 282-283).

Making collections valuable and re-usable for researchers does not have to involve a large-scale effort to build platforms and maintenance-heavy metadata structures for search. Researchers are eager to become involved. They are eager to use the collections that exist and to create their own. One of the primary obstacles to the involvement of researchers in early phases of web archiving projects, though, is a lack of user-friendly **interfaces**. While the tools for capturing and documenting websites are now in place, there are still not sound, intuitive interfaces for interacting with web archives, particularly at the scale of the larger archives. Currently, in order to access an archived version of a website in most collections, users must know the URL of that site. Searchability of web archives is still minimal. If a site no longer exists it is therefore buried, unless the user remembers the site's URL or finds it via an archived hyperlink from another site. The scale of web archives alone presents challenges for providing usable and intuitive interfaces and the temporal and versioning aspects of web archives compound these challenges further.

Ultimately, and fundamentally, there is an epistemological conundrum about ***what constitutes a document*** in a web archive. This conundrum is at the heart of the disconnect in understanding access points across collections. This is a fundamental and persistent discussion in web archiving. Web archiving is a creative process. For each "archived object" we have an impression that it is an approximate representation of what was on the live web. We cannot verify its veracity with the live web. As web technology advances, the notion of the "live web" becomes less and less static - web objects are served differently to different people. Our archived impressions are often incomplete. At times they are loose representations of the objects we wish to capture. At worst, they are snapshots of one instantiation of a dynamic object that may look, in detail on the live web, very different to the many individual users viewing simultaneously. As web historiography develops as a field, it will no doubt develop different methodological approaches to dealing with this epistemological problem.

The problems described above are only a sample. The challenge to web archivists and those building tools to support their use is to build sustainable systems that can weather the coming epistemological rifts in methodology that will arise as the field grows. This epistemological conundrum begins at the earliest stages in the web archiving planning process and continues through to research, and takes hold in the subsequent re-use of previously collected archives. It is this epistemological conundrum that makes many current web archives difficult to re-use.

There are so many different valuable research-oriented approaches to an archive. These approaches, or methods of search and retrieval, are often reduced to tools that represent the few most basic methods (e.g., full-text search without lexical indexing, or specific item search and retrieval based on strict metadata points). Other richer and more powerful search strategies focus less on searching, but rather more on temporary sorting. These methods are experimented with largely in research settings where researchers are working alongside archivists and librarians to build robust collections. As collections are being built, and as researchers are using them, they can add value themselves. Their additions, in turn, make the collections valuable and re-usable for future users. Each new slice through the collection by each new researcher adds to the robustness and re-usability of the collection. Each new way of searching through the data may not be valuable in and of itself to the next researcher who uses the collection, but it may spark interest and creativity.

While the IIPC toolset mentioned above is becoming more heavily embedded into the web archiving practices of institutions, the creation of web archives by individual researchers and end-users is still an elusive and often *ad hoc* practice. The goal among all involved in web archiving should be to turn existing, institution-level technology and resources into accessible and stable services that any user in any discipline can share, adapt, and repurpose. This statement is made with the current culture of personalization, Web 2.0 and 3.0 technologies, and the self-directed and democratic characteristic of web culture itself in mind. The focus on the ability to re-use, repurpose, and personalize research resources mirrors this trend in web culture, and shifts focus to the users' role in developing, not only making use of, humanities and social science research resources for the web. Perhaps users are a valuable resource for web archiving documentation that is yet to be tapped.

Web archives challenge: Knowing the users

One of the big challenges for the organizations who host web archive collections is that it is difficult for them to know how, or even if, their collections are being used. According to Ed Pinsent of the University of London Computer Centre, *"Not much is known about the users of the JISC web archive. The public do not feed back to the JISC or to ULCC as to what use they make of the collections. The only evidence we have is statistical evidence, generated from the log files by the British Library. But this simply records visits to the UK Web Archive and doesn't tell us anything about who these people are, why they are visiting, what they expect to find when they get there, what they take away with them, or whether they have experienced any degree of satisfaction."* One possible approach is to apply impact tools, such as the JISC-funded *Toolkit for the Impact of Digitised Scholarly Resources* (TIDSR) to web archives, just as they have been to other types of digital collections.

Copyright also remains an obstacle at several levels. In terms of access to archives, in some cases researchers have to go to the library building where the web archive is housed to consult the resource and a result of copy right issues. This obviously makes accessing the electronic resource inconvenient for researchers not located near the archive. In addition, copyright issues regarding harvesting potentially copyrighted content into a new web archive can be difficult to navigate, and the legal issues are not at all clear in this area (Knutson, 2009; Patel, 2007). Also international differences in copyright can stand in the way of international research collaborations and projects. These issues are important to clarify so that researchers and institutions will have greater confidence that their collection building and research can be carried out without infringing the rights of others.

The role of the user

Too little is known about users' behaviours in relation to web archives. Most archiving institutions therefore rely on semi-hypothetical use cases to refine and expand their usability and interfaces. One particularly detailed study was conducted at the National Library of the Netherlands (Ras & van Bussel, 2007). This structured experiment, run similarly to a task-oriented usability study, evaluated user comfort level with search and access tools and attempted to determine user satisfaction with archive contents. Several use-scenarios were posited. Few native users have been studied to date, and reports of these studies remain unpublished works-in-progress. We do not have much to draw on when speculating about users in web archives. However, those who are developing their own web archives for directed and narrow research purposes can provide some insight about how they use their archives to produce knowledge in their field.

Web archives challenge: Chickens and eggs

"We tell them what's possible and we want them to tell us what's useful" – Helen Hockx-Yu, The British Library

From the perspective of libraries and large archiving efforts, working with users presents a "chicken and egg" scenario. Usable web archives are just emerging, such as the one released by the British Library in February 2010, and institutions are just now beginning to understand what is possible. Researchers are being asked how they might like to use web archives, but until recently have not known what is possible. Several user-focused initiatives are being led by institutions such as the British Library and the European Archive, and the results of these studies will be pivotal in understanding what will come next.

Despite the reciprocal relationship between the development of research in the humanities and social sciences, there is tension between archiving practices used by researchers and their subsequent access requirements and archiving practices and perceived access requirements in heritage institutions. Each recognizes value in archiving artefacts from the web, but each has followed different paths to develop web archiving practices with a special focus on characteristics most relevant to their immediate environment. More and more, each community is beginning to understand the particular sets of expertise each community can offer to the cause; and members from each community are beginning to understand the value in partnering to achieve the shared goal of stabilizing and preserving artefacts from the web. Ultimately one aim in these efforts is to develop or identify key elements to support the emergence of an infrastructure for web archiving activities for research in the humanities and social sciences.

Researchers, technology developers, and cultural heritage institutions need to work together in order to build this infrastructure with an acute awareness of preservation, accessibility, and interpretation in all their different permutations in the diverse sets of practice. Keeping a diverse set of users in mind, preservation, accessibility, and interpretation can come to be more inclusive and representative of expert and lay-expert views together. To date, most institutions actively archiving web objects focus on some limiting definition to bound, or stabilize, web objects as documents, and place emphasis on an efficient system for generating metadata to enable smooth transitions between archived web objects and other documents. This is highly influenced by traditional library practices. However, the ephemeral and dynamic nature of web objects questions traditional notions of the document. The unclear definitions of web objects lends itself to experimentation with practices in documentation, notably the inclusion of broad annotating activity by diverse users to describe web artefacts and add value to archives for researchers in the humanities, sciences, and social sciences.

Page is intentionally blank

Recommendations

To draw together existing web archiving technologies into an infrastructure that will support e-research and e-heritage, we must foster community and create an abundance of tools and resources that are usable by a variety of users.

Building Community

- Web archiving resources remain largely inaccessible; **the creation of communities that increase the accessibility and usability of web archiving tools should be encouraged** so researchers and librarians can have a common space to share best practices and develop standards.

- Researchers and librarians are often re-building, re-stabilizing, and re-conceptualizing web archiving for each new project undertaken; **sharing tools and sharing resulting web archives for research should become the norm** for both researchers and librarians. These shared resources should enable participants to share archives in a flexible way that meets both institutional missions and individual research needs. The idea of **virtual collections made up by on-demand integration of information from multiple physical collections** would allow users to create thematic collections with much less effort than at present.

- Contributions are being made on a practice-level, a structural level, and a theoretical-conceptual level, but are disconnected in the scholarly literature and professional practice communities; **new approaches should enable connections across disciplines and professions** to encourage web archiving to grow as a flexible field.

- **Privacy and property issues should be made more understandable** in the web archiving space. Many people working in e-research and e-heritage are limited in their use of tools, sharing of practices, and sharing of results due to international law, institutional missions, publication restrictions, and often individual personal preferences in protecting data and methods. Much more powerful tools (based, e.g. on Digital Rights Management technologies) are needed to allow archivists to collect, and users to navigate ethically and legally through these minefields, and to publish with some confidence that they will not run into future liabilities.

- International collaboration remains an important, albeit costly, element to the continued development of tools, resources and standards. In parallel with these continuing international approaches, **local instances off these collaborative outputs need to be created that can feed back into community meta-collections in order to maximize consortial efforts.** The development of such tools, as exemplified by the *Archives Hub*[35], will help avoid duplication of collection efforts and serve to give users a much richer overview of what content may be available, and where.

Building Tools & Resources

- In balancing between the top-down needs of institutions and the bottom-up needs of researchers, there need to be **two related streams of support: one for infrastructure and one for individual archiving**. Crucial to this two-pronged approach, however, is **building a way to connect the two**.

- Technical obstacles are keeping many researchers and librarians out of the emerging web archiving community. **Tools should be both sharable and easy for researchers and librarians to**

[35] http://archiveshub.ac.uk/

implement. Tested solutions to struggles in technology should be easy to find and execute. Usability in installation and use should be a primary concern in future tool development in order to attract more researchers to working with web archives.

- Current web archiving efforts rely heavily on the same set of existing tools, but few of these are specifically focused on extracting data from archives in a manner that enables serious research. **Efforts should be made to diversify the development of tools and interfaces beyond preservation and into use.** These tools, as mentioned above, should be shared widely as a normal practice. Ideally, such tools, should aim to blur the distinction between live and archived content, and also allow much more powerful visualisations of the structure of complex collections, and their changes over time.

- An approach based on modern software engineering practices (e.g. the establishment of collections of Web services or other programmatic interfaces) would allow the current, rather monolithic tools to be replaced by **collections of standardised building blocks whose activities could be orchestrated by workflow tools**.

- Researchers and librarians struggle to use the archived web in research and heritage because there are currently so few ways to parse the information gathered in a crawl; it should become commonplace for researchers in varied fields to have **tools to execute query searches over multiple web archives** to find themes in content that go beyond the results provided in a full-text and 'presence or absence' search.

- Standards for metadata vary by researcher, field, and tools; it should become commonplace to call up a **typology, or vocabulary, of metadata particular to the line of inquiry** that inspired the original query. Metadata should be relational and movable for the needs of the audience at hand. The development of new metadata standards outside the library community, such as the Resource Description Framework (RDF)[36] and Linked Data[37] conventions, point out new ways in which rich and flexible metadata can be used not only for retrieval but also for linking together documents and data sets from different sources, in different formats.

- **Development of standards, protocols, and methods of quality control will help to make web archives more interoperable.** However, the diverse needs of researchers need to be taken into account, so standards must be built that have the flexibility to accommodate innovative uses.

- **For these archives to find value in the world or research, they need to have multiple access points: administrative, descriptive, and contextual.** Administrative access allows for structural integrity, descriptive access allows one to understand the catalogue of contents in an archive, and contextual access places the artefacts within the archive in a thickly described and purposeful context.

- While considerable effort has been put into developing data archives, there has been considerably less commitment to building places to store and share web archives. **Resources need to be developed that allow researchers to deposit and publish their web archives** that are searchable, with organized metadata, and with transparency in the collection criteria, period of capture, and other technical details so that researchers will know what they are dealing with when accessing and re-using the web archives. The adoption of cloud storage technologies may allow the stretched resources of the Web archive community greater economies of scale, leading to an eventual change from "collecting the needles" (assuming that archivists know

[36] See, e.g. http://en.wikipedia.org/wiki/Resource_Description_Framework
[37] http://www.linkeddata.org

ahead of time what needles their users may interested in) to "collecting the haystack," thereby giving much more freedom to the users to ask unanticipated questions and navigate in unanticipated ways.

Building Practices

- **Web archiving practices need to be integrated into the daily practices of cultural institutions.** Libraries have existing policies and practices for collection development that can and should be expanded to encompass web-based materials.

- To understand the possibilities for research uses of web archives, researchers need to have some understanding of how websites are built and how they behave. **Basic training in the area of web content design can lead to a better understanding of how to capture, archive, use, and interpret content from websites.** If they need to make important decisions based on what is stored in a web archive (for example, a lawyer trying to prove a web page contained a certain image on a certain date), they are certainly going to need to be trained about the basics of HTML, web browsers, CSS, JavaScript, web crawling, and possibly other factors. Or they are going to need an intermediary who can explain to them what they need to know in layman's terms.

- **The possibilities of web archives should be communicated to a much broader research community.** A number of examples of potential uses are given below.

- There need to be **better resources for researchers to be able to match available tools to their research needs**. It is currently too difficult to find and understand which archiving software is going to be useful and user friendly for any given practical use.

- Postgraduate training is an excellent way to engage new researchers with new methods and objects of research. **Funding students to look at questions which require the use of web archives, and providing them with the skills to help create the next generation of tools**, has the potential for enabling considerable growth in web archiving for research and for encouraging creative uses of web archives.

- **Support for experimentation in practices is vital at these early stages as the field is still being defined.** Creative new uses may emerge from unexpected quarters, and providing support for these unexpected innovations is crucial.

- **Mentorship of new researchers is necessary to instil the importance of archiving the materials one studies as one studies them.** We need to encourage our undergraduate, post-graduate, and post-doctoral researchers to follow best practices in archiving the web materials they are studying, to build these practices, and also develop the resources that will be available to researchers for further study.

- Funding bodies such as JISC are increasingly recommending that holders of digital collections measure the impact those collections have on various audiences. **Using methods such as those in the JISC-funded *Toolkit for the Impact of Digitised Scholarly Resources* (TIDSR)[38] to measure and enhance the impact of collections of web archives is good practice.**

[38] http://microsites.oii.ox.ac.uk/tidsr/. Members of the project team creating this report were also responsible for developing the TIDSR resource. Other approaches to understanding audiences and enhancing impact would also be appropriate.

Sample potential uses of web archives

There are many potential uses of web archives. To get researchers thinking about the possibilities of web archives, the following ideas represent examples of the types of questions that either could be answered with current tools and methods, or that could be answered with the development of new tools and methods. Of course, this list is suggestive, not exhaustive; many other areas are possible.

Web archives challenge: Imagining the uses

Niels Brügger, an Associate Professor in the Department of Information and Media Studies at Aarhus University, Denmark, was interviewed and discussed how researchers need to imagine the potential uses of web archives:

I guess I would like to see as many people as possible doing history using archiving stuff. To use this kind of material. Have people using it, asking questions of the archive, developing a little what they can do. In one of my texts, I distinguish between five strata that you can focus on in the web. There's the web as such, then the web sphere (clusters of web sites), you can have a web site, a web page, and the web element. And I would like to see studies in all these strata in a way. I am not advocating that we should only do web site research - they are all important and they are all context of each other. I would like every historical study as possible on all these 5 strata. For example, can one imagine, as Kirsten [Foot] and Steve [Schneider] do the history of a web sphere – that's what they do with their presidential elections. Web sphere analysis. Web site history/analysis is what I try to do web pages, that could be, for instance, we heard Megan Sapnar talk about. The design. Web elements - there was a person at a recent meeting who did not give a presentation, but she is working with ads on the web. Banner ads - that history. That would be the history of the element. I hope that people start doing all these things.

If you want to study the web sphere, the links are crucial. And the web site, the outgoing links might be important. Maybe the targets aren't important, but you want to know that it was a link. Studying pages, there you probably find it necessary to have all the elements on the page. I think that would be important for an archive. And the elements, and again, if you study streaming media, the use of video throughout the history of the web, it is important that the archive have those elements. So each of the strata might pose different demands.

Humanities Scholars

There are many sites on the web covering historical topics. Take the two World Wars, for instance: many sites contain personal testimony and copies of original sources such as photographs, letters and official documentation (Meyer, Carpenter, & Middleton, 2009). It may be that members of the public who might not think to approach an archive or library with their own story or personal mementos would be more likely to mount details or copies of their mementos online, which has happened with the *Great War Archive* project at Oxford[39]. Often people have responded to sites which invite those who lived through these events to contribute their memories, and people may be more willing to do so in the privacy of their homes via the internet or through a local event. One of the attractions of these sites to historians, therefore, might be that they offer previously unavailable or untapped primary sources. Other humanities scholars such as those interested in the web as corpus for linguistics are natural potential users of web archives (Hundt, Nesselhauf, & Biewer, 2007; Kilgarriff & Grefenstette, 2003).

Sample questions include:

- How many photographic sources are available on the web for a particular historical event or time period? If places are tagged in these photos, is it possible to reconstruct a virtual panorama of the place or time in question?
- How many personal reminiscences are available across different websites? Do the same people, events, and places in these reminiscences occur in different accounts? When were the reminiscences written, by whom, and for whom? Tools to find, analyse, and view

[39] http://www.oucs.ox.ac.uk/ww1lit/gwa

related documents that refer the same people, places and events would greatly expand possibilities here, such as those being developed in the *Cultures of Knowledge*[40] project.

- Do alternative sources and accounts that are on the web challenge the current historiography? To what extent have these sources been overlooked by traditional historians? Have the kinds of historical sources and documents available on the web changed over time? Does this tell us anything about either history, or about the practices of historians and those members of the public interested in history?

- Are there topics that are of broad interest to the amateur historians and the public that appear frequently on the web, but are largely absent from the traditional historical discourse? Have amateurs developed interesting areas, or found novel ways to present historical information? Are the documents on the web any more or less reliable that other sources?

- Using the huge amount of language available on the web, what can we understand about language change? How is written language changing to reflect new technology? What languages are rising or falling in dominance on the web?

Internet Researchers and Social Scientists

Scholars who are interested in the Internet and its impact on society are clear candidates to become users of web archives. Most research in the area of Internet studies has been cross-sectional, based on data collected at a particular point in time. Now that the web has been around for the better part of 20 years, there is a need to start understanding changes over time on the Internet. Some examples of the kinds of questions one might ask using web archives:

- How has the growth of online news varied country by country over time? Given the claims made by some newspapers that the Internet is killing newspapers, is there historical evidence for any relationship between the depth of online content and a newspaper's financial solvency? How does the contribution of online news affect democratic debate?

- Where has discussion of climate change been most active? How has this changed over time? Is it possible to map the geographical spread and the topics covered in the debate to the geography of climate change effects and attempts at mitigation?

- What kind of predictive indicators about future potential financial crises can be uncovered through the retrospective and real-time data mining of the web?

- Using hyperlink analysis of the structure of the web to understand the social processes around topics and events. While some hyperlinking behaviour is formal or institutional (e.g. government agencies linking to one another as authoritative sources of information or providers of services), a lot of hyperlinking activity is more informal, reflecting the grassroots networking of bloggers, NGOs, special interest and advocacy groups. How can changes in linking over time help us to understand the role of informal communication as part of the feedback loops influencing developing issues? What hyperlinking behaviour is exhibited by these actors, and how can this be related to social science models of collective action?

[40] http://www.history.ox.ac.uk/cofk/. While this project is not focused on web archives *per se*, it is developing methods for linking between similar references in letters that would be applicable to a researcher looking for these sorts of links in web documents.

- How has the visibility of topics changed over time? Do websites that fall within certain network clusters during one point in time ever move to different clusters, or do they remain stable? For those that move, what separates them from the more stable parts of the network?

- Can we develop better tools to analyse web archives statistically? How many sites exist on certain topics? How has this changed over time? What languages are the pages in? Are there clusters around which pages are created, or have they grown steadily over time? Are certain topics more interlinked than others? Can websites be divided into categories that we can uncover using cluster analysis? Can we compare sites by statistics such as the average size of the website in different categories, average number of links, amount of non-textual data (photographs, images, etc), age of content between updates, frequency of updates, type of interface (static versus dynamic, for instance).

- Can we visualize web archive data using methods such as tag clouds of the website titles or keywords or of all the textual content on the website? Can we do linguistic analysis of the terms and words used, and sub-divide the sites into different clusters linguistically?

- With regard to user creation of content, much of the hype around the web, particularly web 2.0, is that users are creating more and more content. This shift from the passive consumption of media content about the world to active participation in the generation of content is clearly happening in areas such as the creation of YouTube videos. Can we measure anything about this non-professional content creation? For instance, what proportions of the collections reside in different domains (.edu, .ac, commercial domain, yahoo website, etc.)? Can we determine which kinds are more likely professional versus amateur creators? If so, can we distinguish between them using the measures in the sections above (links, types of data, age of content, age of site, size of site, etc.)?

Many other questions are possible, as these are just a few to get people thinking about the possibilities for web archives as research objects.

Conclusion

Building community and tools with the features listed above will result in a shift in perspective in e-research and e-heritage that:

- Recognizes and enables the reciprocal relationship between e-research and e-heritage on and about the web;

- Fosters historical and heritage work as well as contemporary research on and about the web in the humanities, sciences, and social sciences; and

- Establishes a domain of distributed repositories, services, and expertise.

Participants in web archiving have expressed the need for multiple and varied access points to the same archived web resources. Therefore, focusing on the creation of access points that are suitable for different disciplines who are using the same primary resources - can build interdisciplinary communities that cut across fields with shared resources and common methods.

A participatory, inclusive and representative knowledge ecology can achieve what current knowledge management practices have failed to do – create an inclusive knowledge ecology where access means readability, retrievability, connecting disparate and closely related information, and enabling connections between users in order to make meaning that can be used to create new

knowledge but also support preservation. Social, community-built tools provide viable alternatives to authoritative systems that derive their management from strict process, workflow, security and control and can make user-driven meaning-making part of the process of accessibility. Hierarchical and ontological information management cannot include the deep contextual and cultural usage meanings that might easily place one object in multiple categories. The restrictions that arise from authoritative management of knowledge can be avoided with the participatory, inclusive and representative knowledge ecology that is fostered by social, community tools, although an approach that is too decentralized runs the risk of having a chaotic approach to standards, or no standards at all. Or, as Julien Masanès of the European Archive suggested when interviewed, *"what we need is a CERN for web archives."*

Page is intentionally blank

Appendix A: Interviews

For this project, we supplemented desk research with 17 interviews with a number of stakeholders in the web archiving community. We are grateful to the following individuals for generously helping us to better understand how archivists and researchers are engaging with web archives.

Niels Brügger
Associate Professor, Department of Information and Media Studies
Aarhus University, Denmark

Richard Davis
Repository Service Manager
University of London Computer Centre, United Kingdom

Katrien Depuydt
Head of the Language Database Department
Institute for Dutch Lexicology, The Netherlands

Kirsten Foot
Associate Professor of Communication
University of Washington, United States of America

Wendy Gogel
WAX Project Manger
Harvard University Library, United States of America

Alison Hill
Curator, Web Archiving, Modern British Collections
The British Library, United Kingdom

Helen Hockx-Yu
Web Archiving Programme Manager
The British Library, United Kingdom

Hanno Lecher
Librarian, China Studies
Leiden University, The Netherlands

Julien Masanès
Director
European Archive, France

Frank McCown
Assistant Professor of Computer Science
Harding University, United Kingdom

Mark Middleton
CEO, Hanzo Archives, United Kingdom

Martin Moyle
Digital Curation Manager
University College London (UCL) Library Services, United Kingdom

Kris Carpenter Negulescu
Director of the Web Archive
Internet Archive, United States of America

Ed Pinsent
Digital Archivist/Project Manager
University of London Computer Centre, United Kingdom

Steve Schneider
Professor & Interim Dean, School of Arts & Sciences
SUNY Institute of Technology, United States of America

René Voorburg
Crawl-engineer & Coordinator of web archiving
Acquisition and Processing Division – E-depot
Koninklijke Bibliotheek, The National Library of the Netherlands, The Netherlands

Max Wilkinson
Datasets Programme Technical Lead
British Library, United Kingdom

References Cited

Adar, E., Teevan, J., Dumais, S. T., & Elsas, J. L. (2009). *The web changes everything: understanding the dynamics of web content.* Paper presented at the Second ACM International Conference on Web Search and Data Mining, Barcelona, Spain.

Alpert, J., & Hajaj, N. (2008, 25 July). We knew the web was big... Retrieved from http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

Anderson, C., & Wolff, M. (2010, September). The Web is Dead. Long Live the Internet. *Wired*.

Arms, W. Y., Adkins, R., Ammen, C., & Hayes, A. (2001). Collecting and preserving the Web: The Minerva Prototype. *RLG Diginews, 5*(2).

Baroni, M., & Bernardini, S. (Eds.). (2006). *WaCky! Working Papers on the Web as Corpus*. Bologna: GEDIT.

Brock, A. (2005). "A belief in humanity is a belief in colored men": Using culture to span the digital divide. *Journal of Computer-Mediated Communication, 11*(1), article 17.

Brügger, N. (2005). *Archiving websites: general considerations and strategies*. Århus: Center for Internet-forskning.

Burner, M. (1997). Crawling towards eternity: Building an archive of the world wide web. *Web Techniques Magazine, 2*(5), 37-40.

Cho, J., & Garcia-Molina, H. (2000, 10-14 September). *The evolution of the web and implications for an incremental crawler.* Paper presented at the 26th International Conference on Very Large Databases, Cairo, Egypt.

Chu, S.-C., Leung, L. C., Van Hui, Y., & Cheung, W. (2007). Evolution of e-commerce Web sites: A conceptual framework and a longitudinal study. *Information & Management, 44*(2), 154-164.

Day, M. (2003). *Preserving the fabric of our lives: A survey of web preservation initiatives*. Paper presented at the European Conference on Research and Advanced Technology for Digital Libraries, Trondheim, Norway.

Dougherty, M. (2007). *Archiving the Web: Collection, Documentation, Display and Shifting Knowledge Production Paradigms (Ph.D. thesis).* University of Washington, Seattle.

Dougherty, M. (2008). *Making web archives valuable for researchers: Exploring the state of the art, annotation practices, and possibilities for progress*. VKS Working Paper. Virtual Knowledge Studio. Maastricht, The Netherlands.

Dougherty, M., Foot, K. A., & Schneider, S. M. (2010). *Ethics in/of Web Archiving*. Paper presented at the Computer Supported Cooperative Work Pre-conference on Revisting Research Ethics in the Facebook Era: Challenges in Emerging CSCW Research, Savannah, GA.

Dougherty, M., Schneider, S. M., & Jones, J. (2010, Forthcoming). Web Historiography and the Emergence of New Archival Forms. In D. W. Park, S. Jones & N. W. Jankowski (Eds.), *The Long History of New Media: Technology, Historiography, and Newness in Context.* New York: Peter Lang Publishing.

Fetterly, D., Manasse, M., Najork, M., & Wiener, J. (2004). A large-scale study of the evolution of web pages. *Software-Practice and Experience, 34*, 213-237.

Foot, K. A., & Schneider, S. M. (2002). Online action in campaign 2000: An exploratory analysis of the US political Web sphere. *Journal of Broadcasting & Electronic Media, 46*(2), 222-244.

Foot, K. A., & Schneider, S. M. (2006). *Web campaigning*. Cambridge, MA: The MIT Press.

Franklin, M. (2005). *Postcolonial Politics, the Internet, and Everday Life: Pacific Traversals Online*. London: Routledge.

Hackett, S., & Parmanto, B. (2005). A longitudinal evaluation of accessibility: Higher education web sites. *Internet Research, 15*(3), 281-294.

Hendler, J., Shadbolt, N., Berners-Lee, T., & Weitzner, D. (2008). Web Science: An Interdisciplinary Approach to Understanding the Web. *Communications of the ACM, 51*(7), 60-69.

Hodge, G. M. (2000). Best Practices for Digital Archiving: An Information Life Cycle Approach. *The Journal of Electronic Publishing, 5*(4).

Hundt, M., Nesselhauf, N., & Biewer, C. (Eds.). (2007). *Corpus linguistics and the web*. Amsterdam: Editions Rodopi B.V.

Innis, H. (1951). *The Bias of Communication*. Toronto: University of Toronto Press.

JISC. (2008). PoWR: The Preservation of Web Resources Handbook. London: JISC.

Johnson, P. (2009). *Fundamentals of Collection Development and Management (Second Edition)*. Chicago, IL: American Library Association.

Kahle, B. (1997). Preserving the Internet. *Scientific American, 276*(3), 82-83.

Kahle, B., Prelinger, R., & Jackson, M. (2001). Public access to digital material. *D-Lib Magazine, 7*(4).
Kelly, K. (2006, 14 May). Scan this book! *New York Times Magazine*.
Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics, 29*(3), 333-347.
Knutson, A. N. (2009). Proceed with Caution: How Digital Archives Have Been Left in the Dark. *Berkeley Technology Law Journal, 24*(1), 437-473.
Koehler, W. (2004). A longitudinal study of Web pages continued: a consideration of document persistence. *Information Research, 9*(2).
Levinson, P. (1997). *The Soft Edge: Natural History and Future of the Information Revolution*. New York: Routledge.
Lyman, P., & Kahle, B. (1998). Archiving Digital Cultural Artifacts: Organizing an agenda for action. *D-Lib Magazine, July/August*.
Lyman, P., & Varian, H. R. (2003). How Much Information 2003? Retrieved 18 August, 2010, from http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/
Masanès, J. (2002). Towards Continuous Web Archiving: First results and an agenda for the future. *D-Lib Magazine, 8*(12).
Masanès, J. (2005). Web Archiving Methods and Approaches: A comparative study. *Library Trends, 54*(1), 72-90.
Masanès, J. (2006). *Web archiving*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.
McEnery, T., & Wilson, A. (2001). *Corpus linguistics*. Edinburgh: Edinburgh University Press.
Meyer, E. T., Carpenter, K., & Middleton, M. (2009). World Wide Web of Humanities: Final Report to JISC. Online: http://www.jisc.ac.uk/media/documents/programmes/digitisation/humanitiesfinalreport.pdf. London: JISC.
Morville, P. (2005). *Ambient Findability: What We Find Changes Who We Become*. Sebastopol, CA: O'Reilly Media, Inc.
Murphy, J., Hashim, N. H., & O'Connor, P. (2008). Take Me Back: Validating the Wayback Machine. *Journal of Computer-Mediated Communication, 13*(1), 60-75.
Patel, K. (2007). Authors v. Internet Archives: The Copyright Infringement Battle over WEB Pages. *Journal of the Patent and Trademark Office Society, 89*, 410-428.

Ras, M., & van Bussel, S. (2007). Web Archiving User Survey. Retrieved from http://www.kb.nl/hrd/dd/dd_projecten/webarchivering/documenten/KB_UserSurvey_Webarchive_EN.pdf
Raymond, M. (2010, 28 April). The Library and Twitter: An FAQ. Retrieved from http://blogs.loc.gov/loc/2010/04/the-library-and-twitter-an-faq/
Schneider, S. M., & Foot, K. A. (2002). Online Structure for Political Action: Exploring Presidential Campaign Web Sites from the 2000 American Election. *Javnost-The Public, 9*(2).
Schneider, S. M., & Foot, K. A. (2004). The web as an object of study. *new media & society, 6*(1), 114-122.
Schneider, S. M., & Foot, K. A. (2005). Web Sphere Analysis: An Approach to Studying Online Action. In C. Hine (Ed.), *Virtual Methods: Issues in Social Research on the Internet* (pp. 157-170). Oxford: Berg.
Schneider, S. M., & Foot, K. A. (2010). Object Oriented Web Historiography. In N. Brügger (Ed.), *Web History*. New York: Peter Lang Publishing.
Simonite, T. (2010). A Search Service that Can Peer into the Future: A Yahoo Research tool mines news archives for meaning--illuminating past, present, and even future events. *Technology Review*. Retrieved from http://www.technologyreview.com/computing/26113/
Taycher, L. (2010, 05 August). Books of the world, stand up and be counted! All 129,864,880 of you. Retrieved from http://booksearch.blogspot.com/2010/08/books-of-world-stand-up-and-be-counted.html
Taylor, M. K., & Hudson, D. (2000). "Linkrot" and the usefulness of Web site bibliographies. *Reference & User Services Quarterly, 39*(3), 273-276.
Thelwall, M., & Vaughan, L. (2004). A fair history of the Web? Examining country balance in the Internet Archive. *Library & Information Science Research, 26*(2), 162-176.
van den Heuvel, C. (2009). Web Archiving in Research and Historical Global Collaboratories. In N. Brügger (Ed.), *Web History* (pp. 279-303). New York: Peter Lang Publishing.
Veronin, M. A. (2002). Where Are They Now? A Case Study of Health-related Web Site Attrition. *Journal of Medical Internet Research, 4*(2).
Weinreich, H., Obendorf, H., Herder, E., & Mayer, M. (2008). Not quite the average: An empirical study of Web use. *ACM Transactions on the Web, 2*(1), 1-31. doi: http://doi.acm.org/10.1145/1326561.1326566