

Community, Tools, and Practices in Web Archiving: The state of the art in relation to social science and humanities research needs

Meghan Dougherty
Loyola University Chicago, School of Communication
820 N. Michigan Ave
Chicago, IL 60611
(312) 915-8834
mdougherty@luc.edu

Eric T. Meyer
University of Oxford, Oxford Internet Institute
1 St Giles
Oxford OX1 3JS
United Kingdom
+44 (0)1865 287218
eric.meyer@oii.ox.ac.uk

This is the peer reviewed version of the following article: Dougherty, M., Meyer, E.T. (2014). Community, Tools, and Practices in Web Archiving: The state of the art in relation to social science and humanities research needs. *Journal of the American Society of Information Science & Technology* 65(11): 2195-2209., which has been published in final form at <http://dx.doi.org/10.1002/asi.23099>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

Abstract

The Web encourages the constant creation and distribution of large amounts of information; it is also a valuable resource for understanding human behavior and communication in the late 20th and early 21st centuries. To take full advantage of the Web as a research resource that extends beyond the consideration of snapshots of the present, however, it is necessary to begin to take web archiving much more seriously as an important element of any research program involving web resources. The ephemeral character of the Web requires that researchers take pro-active steps in the present to enable future analysis. Efforts to archive the Web or portions thereof have been developed around the world, but these efforts have not yet provided reliable and scalable solutions. This article summarizes the current state of web archiving in relationship to researchers and research needs. Interviews with researchers, archivists, and technologists identify the differences in purpose, scope, and scale of current web archiving practice, and the professional tensions that arise given these differences. Findings outline the challenges that still face researchers who wish to engage seriously with web content as an object of research, and archivists who must strike a balance between a range of user needs.

Introduction

As more of the daily activities of life are carried out on the Web, the Web pages, sites, and applications that enable those activities and record traces of our communications are increasingly becoming a key medium of record for modern society. Information distributed on the Web encompasses a vast array of the activities and artifacts of humanity (Kelly, 2006). Because the Web has enabled the constant creation and distribution of unprecedented quantities of information, it is a key resource for understanding human behavior and communication in the late 20th and early 21st centuries. At the same time, however, the content and structure of the Web is constantly in flux as it is updated, replaced, moved, recombined, and deleted. While this constant change is part of the dynamism that has helped to make the Web so successful, the unintended consequence of dynamism is that we are relying on a cultural webscape that exists constantly in the present, limiting possibilities for retrospective analysis. This ephemeral character of the Web requires that researchers who wish to do analysis in the future, or to enable others to do so, must take pro-active steps in the present. In light of this need, various efforts to archive the Web or portions thereof have been developed around the world, but these efforts so far have not yet provided reliable methodological solutions for researchers who wish to use archived web materials. To take full advantage of the power of the Web for understanding society, it is necessary to begin to take web archiving much more seriously as an important element of any research program involving web resources (Schneider, Foot, & Wouters, 2009).

This article summarizes the state of web archiving in relationship to research needs using interviews with researchers, archivists, and technologists to identify obstacles to building web archiving projects, the conceptual and methodological origins of those obstacles, and suggests possible solutions. This study focuses in particular on the needs of individual researchers with regard to archiving web content for retrospective study, and identifies the differences in purpose, scope, and scale of current web archiving practice, and the professional tensions that arise given these differences. A key theme throughout is the shared challenges that researchers and archivists who wish to engage seriously with web content as an object of research currently face, but approach from different and sometimes conflicting disciplinary viewpoints and practices.

One primary challenge that must be addressed if web archives are to become more serious sources of research material is to confront the current disconnections between stakeholders. There is currently a wide gap between the researchers who need archival datasets to support their studies of online phenomena, and the archivists and other practitioners who have the expertise to build such collections and the tools to manage and access them. There is also a gap between the potential community of researchers who have good reason to engage with creating, using, analyzing and sharing web archives, and the actual (small but growing) community of researchers currently doing so. These researchers, archivists,

technologists and other practitioners each identify the purpose and use of archives differently, and so rarely see eye-to-eye on solutions for web archiving even in terms of solutions to shared obstacles.

This article identifies some of the main reasons for archiving web pages, sites, domains, and other web-based content. It presents an overview of the current diverse practices as they are evident in a variety of inquiry modes, attempts at standardization across those practices, and the obstacles faced by researchers and archivists wanting to archive web content. Finally, through research interviews with technicians, archivists, and researchers who create and use web archives, strategies for moving forward to build awareness and practices across disciplines to result in more robust archival resources for Internet research are suggested. This approach is informed by the premise that “[a]rchivists should promote a symbiotic relationship with researchers, who, after all, have more time to focus undivided attention on details and who often come to a project with some degree of subject expertise” (Kaplan & Mifflin, 2000, pg. 96). The focus though, is on the current state of engagement with web archives – how are researchers and archivists currently making use of web archives and what technical and policy infrastructures do they need to facilitate their work?

The Current State of Web Archiving

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 share knowledge and information. The Web is enabling steep increases in the rate at which textual, visual, and audio information is being produced and shared. In 2008, for instance, Google reported that their systems had found 1 trillion (10^{12}) unique URLs on the Web at once (Alpert & Hajaj, 2008). The sheer quantity of data appearing on the Web represents a rapid expansion in recorded human knowledge, and includes an increasingly comprehensive record of information production and social interaction over time as more activities become web-enabled or web-based. On the one hand, some tend to dismiss this voluminous creative work as a loose record of cultural production, rather than as an ephemeral site on and within which social interaction is enacted (Taylor & Every, 2000). However, when interviewed for this project, Kirsten Foot, faculty member at the University of Washington, Seattle and co-founder of Webarchivist.org, explained the scholarly consequences of the rapid expansion of content online this way:

“...if we don’t capture the online phenomena in at least the same rigor that we archive newspapers and other kinds of artifacts 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 key aspects of the Web” (Foot, interview).

Over the past twenty years, most of the content of the Web has disappeared as it is replaced by new pages and new content. There is a rapid turnover: several studies found that within a given week 35-40% of web pages changed their content (Cho & Garcia-Molina, 2000; Fetterly, Manasse, Najork, & Wiener, 2004), and that this change is even more rapid when considering dynamic pages such as news and social network sites which are built around the idea of nearly constant updating. Other studies have found that 69% of web sites changed when revisited after a day or more (Weinreich, Obendorf, Herder, & Mayer, 2008), and that certain dynamic information is likely to change more frequently than once an hour (Adar, Teevan, Dumais, & Elsas, 2009). While these pages are updated and refreshed continuously, 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).

Even when changes are archived, they are typically kept in private in-house logs. Furthermore, even the idea that all content and content changes are stored somewhere are a key point of contention for those arguing for the dystopian Internet-as-panopticon view in which no post, picture, or other social communicative interaction online is ever truly lost or deleted, even (or possibly especially) when the source or object of that content wishes to exercise their right to be forgotten (Mayer-Schönberger, 2009). Only recently have social networking sites enabled trails of past activity to be visible (e.g., Facebook’s Timeline), and even in those situations the managing site enables users to edit and remove past posts to enable users to maintain control of presentation of self in online social networks. This is excellent for users wishing to exercise more control over their personal web presence, but it can be problematic for researchers examining the evolution of phenomena enacted on the Web.

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 such as HTML5, burgeoning mobile platforms, and the use of executable content such as JavaScript or Flash – all pose new challenges for the Web archive community (Meyer, Thomas, & Schroeder, 2011). Worse still from the archival perspective, much of the Web’s content is increasingly hidden behind forms-based query interfaces where the actual content is held in databases that are inaccessible to crawlers. The development of methods to allow the content of this “deep web” to be collected poses another major challenge. Other, even more fundamental changes, such as the growing pervasiveness of social media sites such as Facebook and Twitter, among many others, point to a potential sharp decline in the relative prominence of the information structures used in the early web (Anderson & Wolff, 2010). In this new

world, there is a risk that open content, protocols and interface behaviors will be replaced by closed systems, content and interactions which are absolutely invisible to traditional archiving practices in what Meyer et al. (2011) have called the ‘apocalypse’ scenario for the future of web archives.

Creating archives of web objects is a complicated problem. Starting in the mid-1990s, researchers began partnering with librarians to solve some of these problems. Technical solutions are offered to different aspects of the larger problem in a range of projects. Warrick (McCown et al., 2006) at Old Dominion University, which uses the *Memento TimeMaps*¹ platform to discover and reconstruct a lost website. Memento and Warrick enable users to identify a particular resource or browse by date and can use existing repositories of old copies to restore lost data in what has been termed Lazy Preservation (McCown, et al., 2008). McCown and Nelson have used this work of lazily preserving websites and rebuilding them to build a framework for describing extant web repositories and the status of web resources in them, and offer an API to “serve as a foundation for future web repository interfaces” (2009).

Some of these technical solutions have been used to create archives of web objects that could be queried to draw generalizations about a variety of topics in the humanities and social science. Research analyzing content of web archives range from studies about politics enacted on the Web (Foot & Schneider, 2006; Kluver, Jankowski, Foot & Schneider, 2007), to explorations of the Web presence of different cultures (Franklin, 2005), to linguistic studies (McEnery & Wilson, 2001), to examinations of public reaction to world political events on Twitter (Starbird & Palen, 2012). These types of inquiry have contributed to shaping the descriptive, methodological, and theoretical bases of scholarship centered on web archives.

As web archives have become more accessible and more widely known, researchers and librarians worldwide have begun to investigate how web archives can become a resource that complements exploration of the live and actively changing web. Approaches to Web archiving tend to fall into three categories: large-scale collections, smaller-scale thematic collections, and idiosyncratic collections (Dougherty, Meyer, Madsen, van den Heuvel, Thomas & Wyatt 2010).

Large-scale projects such as the Internet Archive and the Internet Memory Foundation have parallel missions to make their collections accessible to the widest audience possible. To date, their efforts have been primarily focused on providing timestamped snapshots 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 Internet Memeory Foundation, too, is focused on building tools that allow researchers to engage with their archives, for example to run analytics, perform linguistic analysis, or leverage social media for collection development (Masanès, 2013). A number of studies have established the Internet Archive as a valuable resource for research in the social sciences to estimate 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). Hackett & Parmanto (2005) used the Internet Archive to analyze changes in website design in response to technological advances over time. Efforts along these lines include *Memento TimeMaps*, which adds a time dimension to the HTTP protocol to better integrate the current and past web, and the *Yahoo Time Explorer*,² which is being developed to build timelines from searches in news archives. This previous work has clearly established the utility of data from the Internet Archive as a source of research data (Brock, 2005; Thelwall & Vaughan, 2004; Veronin, 2002; Murphy, Hashim & O'Connor, 2008; Chu, Leung, Van Hui & Cheung, 2007; Hackett & Parmanto, 2005). 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).

Other web archiving approaches are selective, thematic, deposit-based or a combination of these approaches. Selective approaches identify artifacts 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. This type of selection at the harvesting level is employed by PANDORA³ (Preserving and Accessing Networked Documentary Resources of Australia), 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.⁴ Deposit-based projects, such as projects at the National Library of the Netherlands⁵ (Koninklijke Bibliotheek), rely on voluntary deposits. 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 artifacts within a larger broader sweep of content. The Digital Archives for Chinese Studies⁶ (DACHS) with branches at the University of Heidelberg and Leiden University, and Virtual Remote Control⁷ (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."⁸ Several of Harvard University's libraries 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*⁹ and the *Constitutional Revision in Japan Research Project*.¹⁰ 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.

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 research project guides the criteria for selecting objects for inclusion. Categorization follows coding strategies informed by prior inquiry into the field and developed to address certain concepts to be tested in the study at hand. These collections are limited in size and scope. 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 only 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. Furthermore, because these archives are built on a shoestring budget by a researcher who may have little to no understanding of archiving procedures, and have no real technological infrastructure to rely on, they are often inaccessible to others, residing on the hard drives of individual researchers.

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, among other dimensions of traditional archival principle.

Advocates of web archiving draw on methods in digital cultural heritage to manage the quantity and variety of web archival data available, in the hopes of advancing the potential for studying new genres such as blogs, microblogs, and social media sites. 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). Skeptics, however, have questioned the trustworthiness of archives collected by researchers, arguing that control over sources and long-term stability of such collections should be better defined (Brügger, 2005).

Many debates about the potential uses of web archives still remain at both a theoretical and practical level, but 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 begun 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, 2010; Schneider & Foot, 2010), the ethical and legal impacts of

saving artifacts 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 and other search engine queries to find artifacts from a web sphere to capture and archive (Schneider & Foot, 2004), and expert derived sets of artifacts 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 (e.g., see the papers in Baroni & Bernardini, 2006).

While this work has provided interesting tools and new insights, there have thus far been only isolated successes by organizations interested in making an infrastructure that melds the dynamic web with the stability and control of traditional methods in archive research available to the larger research and heritage community. One such example is the PANDORA¹¹ web archive in Australia, which is a selective archive involving a number of Australian institutions.¹² Another such model is the California Digital Library's Web Archiving Service.¹³ While many of the institutions using the CDL/WAS service are located in California, there are also examples of other institutions utilizing the service for institutional web archiving, such as the Bentley Historical Library at the University of Michigan (Shallcross, 2011). The University of Michigan, as described by Shallcross, follows an archival model that is relatively expertise-intensive, relying heavily on professional archivists to make decisions about which portions of the University's web presence to preserve (Deromedi & Shallcross, 2011). This approach has been demonstrated to work at the institutional level, but once the topic of interest moves to Web-scale, these techniques become difficult or impossible to use at scale.

Methods

To understand the gap that has developed during this critical stage in web archiving practice, we sought out responses from key researchers, archivists, and information technologists active in the field. Semi-structured in-depth interviews were conducted with seventeen professionals in the U.S., U.K., and Europe during 2008 and 2010. Since the Web archiving community is relatively small, we used a purposive sampling method to identify some of the key experts in web archiving.

Each of the six interviewers in the study contributed a list of possible interviewees for inclusion in the study. These names were generated from interviewers' familiarity with active members in their particular sub-field of web archiving. The subfields within which the interviewers work include information science, social science research, humanities research, library and archive information systems, and science and technology studies. We observed common threads that tied together our successes and frustrations, but also observed that each of us framed obstacles differently depending on our sub-field creating gaps in practices. We conducted interviews with others in our sub-fields to learn

more about the gaps we observed in our own experiences, and sought to use the interviews to describe those gaps with more clarity.

The interviewees represented a mix of Internet researchers using web archives, other social science and humanities researchers trying to use web content for research, and archivists and information managers working in different subject areas building web archives for different purposes. Respondents are all prominent figures in the field, authors of well-cited articles in the field, or participants in oft-cited web archiving initiatives. Of the 17 people interviewed, four are researchers in social science and humanities, nine are archivists or librarians working on digital preservation projects at their institutions, and four are technicians or software engineers building tools to support digital preservation (See Appendix A). Several interviewees had cross-over roles, for example, of the nine interviewees who are archivists or librarians working on digital preservation projects at their institutions, all lend some technical building skill to their projects, but those we identify as technicians or engineers described themselves as “Technical Leads” or “Engineers” or “Managers” and “Computer Scientists” of specific projects. Other cross-over roles emerged in the process of interviewing where researchers described their ad-hoc experimentations with web archiving to support their research acting both as a researcher and technician. The categories and cross-overs of such categories are important to illustrate the differences in approaches to framing and solving web archiving problems.

We supplemented the original purposive sample with snowball techniques: each respondent was asked to suggest names of additional possible interviewees that fit this description. Because this data was collected in an exploratory study assessing the state of a field, dimensions for consideration, and suggestions for future development, a small number of targeted interviews were sufficient to reveal specialized (and often tacit) knowledge from within different groups engaged in web archiving. Interviews were conducted in person, over the phone, and via email. This was an international collaboration, and interviews took place across a number of time zones. Each interviewer was left to negotiate the most convenient medium for themselves and their respondents.

Interviews conducted in 2008 were open-ended, and those conducted in 2010 were semi-structured. Each interview aimed to solicit opinions, ideas and reflection from the respondent based specifically on his or her own personal experience with web archiving. In addition to an unstructured portion of the interview, the interviewers guided discussion with respondents using a set of common themes. These themes included background information on their field of study, discipline, education, time spent with web archives, and others. Interviewers also probed about the interviewees' experience with the practicalities of working with web archives, and their experience with funding web archiving projects. Interviewers questioned interviewees about the perceived value of web archiving, the current landscape of tools available, and standardization of practices. In addition to these common themes, interviewers had

sets of questions to draw from that were specific to archivists, researchers, or technicians that addressed details of each groups' perspective on web archiving. Questions included in the interview schedule were offered to interviewers as a guide to familiarize themselves the boundaries of discussion described by the study, to offer specific questions to prompt interviewees, and to provide a list of topics and themes to be addressed in the course of conversation with interviewees. Because interviews were with people representing a range of roles involved in web archiving, a singular set of questions would not fully address the breadth of conversation described by the study. Rather, a list of suggested themes, topics, and questions could enable interviewers to gather enough complementary data to make comparisons across different approaches to web archiving, and discuss how shared obstacles may be approached differently. The interviewers were given the flexibility to draw these questions into conversation as they saw fit. Each interviewer has his or her own experience in the field, and they were encouraged to use that experience to help them in the interview process.

Findings

Although the respondents' experiences with web archiving were from a variety of different disciplinary perspectives and professional roles—researcher, archivist, technologist, etc.— they all spoke about obstacles to advancing web archiving, reasons for those obstacles, and potential solutions to overcome those obstacles in different ways. The examples they offered showed the particulars of their approach to web archiving from the point of view of specific stakeholders. Despite these particulars there were many common themes throughout the data identified as foundational problems. For instance, several interviewees spoke about what could be categorized as ontological and epistemological approaches to web artifacts and identified their involvement in both concrete local decisions and more wide-ranging professional debates on how best to integrate archives of such objects into existing collections.

The lack of shared practices, accessible tools, and clear legal and ethical guides were repeatedly named as obstacles to advancing web archiving. Definitions of users and their needs varied widely and were linked to public perception, or more accurately the lack of public perception that web archiving is necessary or of any value. Some spoke of difficulties in collaborations across communities in an effort to make these collaborations more transparent. Solutions to all of these issues were posed whether they represented the imaginings of an interviewee's ideal world or current projects tackling some aspect of a larger problem. Each of these themes—ontological approach, obstacles, users, and solutions— are explored separately, but through each theme the conceptual and methodological differences between stakeholders can be seen as a foundational rift.

On electronic records and the Web

One theme that emerged from the data was from respondents who strongly expressed an ideal vision of archived web objects being served parallel to other types of collections. For some respondents, there seems to be a disconnect between the experimental collection and maintenance processes for some web archive projects compared to practices of more traditional archival documentation strategy for which there are well-established professional principles (Cox, 2000). Recommendations for reinventing archiving to address “why current methods fail for electronic records (Bearman & Hedstrom, 2000, pg. 552)” have generated a heated debate about the nature of electronic records and the necessity for new principles, practices, and methods to meet the needs of preserving evidence of social activity. Bearman and Hedstrom (2000) explain that these issues continue throughout the archival process from surveying, appraisal, accession, and preservation activities through to when records are accessed, examined, and analyzed by users such as researchers. Others reaffirm the ability of traditional archival principles and methods to meet the needs of digital culture (Henry, 2000; Gilliland-Swetland, 2000), and have suggested that Shellenberg’s concerns about “how to meet current challenges on the basis of present practices and resources, not starting over again from scratch” (Henry, 2000, pg. 588) is apropos as we consider preserving evidence of social activities on the Web. Archivists interviewed for this study are aware of these theoretical debates, and must navigate between those debates and practical concerns. Because web archives require technical infrastructural support on both the collecting and the reading ends, they are typically accessed separately from other collections, which have different technical requirements for supporting infrastructure.

“My view is that web archives shouldn't be considered in isolation from other digital collections, but should be used to supplement them and considered part of an integrated whole that includes (for example) digital records, archives, books, articles, images, etc. I look forward to a time when an academic digital archive will serve all related materials to a researcher through a single portal, regardless of the formats in which they were created” (Pinsent, interview).

Archivists and users of archives alike see the potential for digitization of materials to be a great leveler for all archived media. Remediating archived content so different types of resources can share a common element for access (e.g., digitizing content stored in other media such as tape, paper, etc.) could refocus the processes of collection, access and use on content rather than on medium.

Traditional ways of treating materials were driven in part by the affordances of the media themselves. Applying the media ecological affordances of older media to new media archives is misguided and can limit the possibilities of all aspects of identifying, collecting, indexing, retrieving, and rendering archived web objects.

“I also think that there is a tendency to treat websites as though they were library books - written by one author, on a single subject. They are often collected and arranged like titles in a library.

This thinking has affected the way harvesting software works, the way that collections are built, and the way that they are described and rendered” (Pinsent, interview).

Treating web archives separately from other media archives, but using the same traditional archival approach will have long-term effects. The more we isolate access to web archives from other archives, the less attention they will receive, and the less progress will be made. Interviewees noted that traditional archival practices may constrain possibilities for how web objects are surveyed, harvested, described, and rendered. Some organizations are starting to realize this: the British Library, for instance, aims to embed web archiving into their regular library workflows where currently bibliographers spend 5% of their time on selecting and archiving web sites. . How many other organizations make similar decisions remains to be seen.

On Obstacles

Respondents described a number of obstacles to web archiving, which have slowed the advance of web archiving, or set it in a direction for which we cannot accurately gauge the future implications. The obstacles either caused respondents frustration in a current or past web archiving project, have derailed collaborations and funding opportunities, or resulted in re-orienting a project as methods to work around intractable problems were implemented. Respondents described unclear or restrictive legal policies, less-than-transparent collaboration, lack of public support, a lack of best practices, and technology limitations as obstacles.

Legal concerns. Alison Hill (Curator, Web Archiving, Modern British Collections, British Library) sees legal concerns, specifically the legal requirement to get explicit prior permission to archive web objects, as a key obstacle to more widespread creation of web archives in the UK; it should be noted, however, that since the time of the interview, the legal situation in the UK has changed, as noted below. The British Library started web archiving in 2004 with a domain project experiment. They selected 100 domains across the UK. They created a policy whereby they would ask permission from the owners to archive these websites—a policy they still follow. This kind of policy is a tradeoff. It proactively protects rights of content creators whether or not they have signaled the extent to which they reserve copyrights, but limits the scope of what the library can collect. It also precludes wide-scale harvesting of the UK web domain. Although efforts to establish regulations that would allow legal deposit libraries in the UK to archive all UK web content have been implemented in 2013 after being discussed and debated for nearly a decade, the resulting archives are still only accessible to researchers working at a physical terminal located in the deposit library.

Researchers commented that legal considerations are especially concerning in inter-institutional collaborations. These legal concerns can serve as a foundation for direction of development and outcomes

of any given project by determining what objects may or may not be included in a collection, and dictating who can access resulting archives and where and how they may do so. Social researchers and archivists approach these detailed protocols differently. Where an archivist may be bound by legal considerations, among other things, to determine what is included and excluded from a collection, a social researcher must have specific details about inclusion and exclusion criteria so he or she may precisely qualify how findings may be generalized. As Foot explained in her interview, “People from different types of disciplines have different concepts in mind even when we use the same terms and it is important to surface those differences.” She was particular about the definition of what it means to be “systematic” and the level of inclusion and exclusion criteria for collection development. There are different practices in different professional communities and domain expertise. Thoughtful agreement around these issues are increasingly important and, “it is important to really thrash through those [differences] and work out a protocol” (Foot, interview). Benardou, Constantopolous, Dallas, and Gavrilis (2010) called for “a closer focus on requirements stemming from actual information work in scholarship” especially in regard to developing of Web-accessible digital resources of interest to humanists. This call is required for social scientists as well to address their specific methodological and informational needs. Collaboration is needed beyond discussion of the legal policies that can constrain these needs.

Collaboration and inter-institutional partnerships. Collaboration and partnership is a complex issue—one that is essential to the success of large-scale web archiving projects. Various partners 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” (Foot, interview). Researchers are excited that universities and institutions that are taking an interest and experimenting with small scale web archiving projects in different ways, but say that they do not seem to have yet developed any standardized strategy for collecting born-digital materials that meet their research needs either in terms of access, analysis tools, or specific evidence relevant to their research (as of 2009).

Beyond negotiating difficult conversations and hashing out these critical differences across disciplines, there are divisions of labor to negotiate in the process of building and using web archives in research in both small- and large-scale projects.

Steven Schneider, a researcher at SUNYIT and co-founder of Webarchivist.org, suggests a shared services approach. He has identified a set of processes in web archiving as a scholarly method: identification, curation, verification, collection, indexation, categorization, and presentation (Schneider, 2006). Researchers need a high level of control over some processes including identification and categorization, and need less control over the other processes. Collaboration is necessary, and a shared service could allocate control along these lines. However, Schneider explains that this division of labor is difficult:

“It is hard because those who control the technical processes (collection, especially; presentation, less so) tend to be the ones who absorb the initial costs, the most significant costs, and the ones least able to be volunteered because of the technical aspect. So if there was a consortium to share these costs, and pre-pay them for a period of time (1-5 years), we could then recruit scholars to participate in this service” (Schneider, interview).

Public perception. Beyond legal and ontological differences, and the unforeseen consequences of choosing a path to move forward in building a particular web archiving project collaboratively, the most problematic obstacle is the perception that web archiving is not valuable enough to commit resources.

“One of the biggest obstacles in the development of web archiving is that people can’t see the point in doing it” (Pinsent, interview).

Unfortunately in this early stage of development of web archiving, as Kirsten Foot explains, “The need outweighs the resources by orders of magnitude.” She explained that as a researcher and web archivist, she is constantly juggling the tasks of parsing through what is being produced online, making sense of it, and also anticipating what will be important when we look back. She pointed out that perhaps an “institutional mindset shift” is in order where institutions including national libraries, local libraries, and universities collect everything they produce in repositories, and to do this for entire institutions by central policy, not just within specific technologically savvy clusters. A shift is needed in archiving policy for web-based content that mirrors paper-based content. She explained that a key driver for this is for researchers to document the resources on which they are making claims in their research, and working with organizations to create simple tools that enable these forms of documentation and citation. This perception obstacle is exacerbated by common user reactions.

“I think for the most part that many people can interact with the Internet Archive’s Wayback Machine OK, but they really struggle with understanding exactly what they are looking at when they see an archived web page. Most people don’t understand that the image they see may not be the same image that was shown on the Web page on the date specified by the Web archive. Most people don’t understand CSS and how a missing style sheet could seriously change the way a website appears in the browser. And most people don’t understand how a web page may have rendered somewhat (or even dramatically) different in the browser years ago compared to how it renders in a modern browser” (McCown, interview).

Web archives are still specialized collections, and the workings of the Web itself remains a bit of a mystery to many (Mitra, 2011). The social and cultural purpose served by the Web and all of its publishing and interactive elements, to say the least, is not fully appreciated by those outside of academic circles of Information and Communication Technology studies, Digital Humanities, and Internet

Research, among a few other specialized fields. The added layer of inconsistency or unreliability in web archives can cast doubt over a whole research endeavor.

Research community perception. Beyond the public views of web archiving are those of funding institutions and research communities that do not see the purpose of archiving something that is happening now on the Web. Mahmood Enayat described a project that was not possible without building custom software, and faced difficulty in finding funding to build that software due, in part in his opinion, to this misunderstood perception that archiving the Web is not necessary. Enayat was a doctoral student at the Oxford Internet Institute at the time of these interviews, and was also a correspondent for the BBC and an oft-consulted expert on the Iranian blogosphere. During the June 2009 Iranian election, he and a group of other researchers were very interested in archiving the Web based materials related to the elections, and they were interested both in official and unofficial materials.

“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 artifacts online. Nothing like this had ever happened before.

Our idea was simply that we should capture these things. For two reasons. 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.

But we just found that there was no proper solution to do it. We looked at Archive-It (the one from archive.org) but it uses Java and has to be installed on a TomCat server and there aren't many web hosting companies that do that. We also talked to Hanzo and it was just way too expensive. So we realized we would have to build something ourselves and that meant getting funding. But no one was interested, they didn't get why we wanted to archive something that was happening right then.

So we started a wiki, to gather all the links together. But of course lots of the stuff now is gone. The youtube videos, some of them are gone and some of the blogs were by people who are now in jail and those are now taken down” (Enayat, interview).

Many scholars face the obstacles Enayat described in this anecdote: developing inclusion and exclusion criteria for the collection of objects that addresses both archival principles, legal constraints, and research standards for generalization; technical constraints stemming from personal familiarity with and

sometimes general confusion about tools and their requirements, or institutional IT support constraints; and writing compelling funding proposals to the right funding institutions to justify the need for an archival infrastructure to support such research. At the time this research was being planned, objects relevant to the research were volatile—appearing, changing, and disappearing quickly—and so an archival strategy needed to be put in place quickly. For a researcher who is not practiced in archival principles, it is a challenge to identify and set up all the different elements to construct the collection of objects needed to do retrospective study in a systematic way. This respondent created a stopgap solution, but was left unsatisfied watching a social phenomenon rapidly develop, change, and ultimately disappear with little to no evidence left to describe it in a scientific way. The gap between the state of social science on the Web, and the state of tools to conduct historical studies on the Web is vast. Producing ad hoc customized software for a small-scale project can seem like a step backward in progress, when in fact it is necessary to start small so we may work out the problems of other obstacles and overcome them.

Community practice. The best practices for developing and maintaining an archive vary depending on the intended use. When the use is not specified, too often practices appear to be chosen at random. When the use is specified, we apply a fitting set of practices, but as a result we often necessarily rule out other subsequent uses.

“...collections of web archives may not have the sort of completeness or comprehensiveness we would want, but these are ideas which have still not been developed or agreed in the community. At the moment I might take four snapshots a year of a project website. Is that too much, or not enough? How do I know when I’ve succeeded? What do the users expect to find in a collection?”
(Pinsent, interview).

Heleln Hockx-Yu (Web Archiving Program Manager from the British Library) explains that it is cause for concern that the whole web archiving community is dependent upon the same set of tools, which increase risks and do not necessarily encourage alternative approaches. This could be a problem going forward, a point also supported by Pinsent:

“The fact [is] that so many of our approaches to website archiving concentrate on solving technical issues and resolving website behaviors, rather than meeting the needs of researchers”
(Pinsent, interview).

The biggest obstacle is overcoming technical problems, so much so that meeting the nuanced conceptual and methodological needs of researchers or other users cannot even be addressed.

Tool development. On the technical side, one of the primary problems is the lack of research and development on the access end of the archival process. Web archives are generally large, cumbersome, and complicated. Access is not intuitive, and the interfaces we do have “...are not sound, intuitive

interfaces for interacting with web archives, particularly with the scale of the archives” (Carpenter-Negulescu, interview).

Alison Hill sees the technologies of the Web advancing faster than the archiving technologies, leaving archiving perpetually one step behind the development of the Web. This sentiment is echoed by many, especially those with a broad scope to their collecting.

“It is going to be more and more difficult to collect and particularly re-render content on the Web. YouTube, for example, changes the parameters used to embed video every 4-6 weeks. Flash based sites are equally problematic. There is going to continue to be huge innovation in the way people publish to the Web. Should they archive, for example, the iPhone's view of the Web? If so, why stop there, what about other mobile devices? The social web is also a huge problem -- the fact that my experience of the Web is different than yours. What about the fact that someone's diary is now being recorded in Facebook? Should they be attempting to capture individual views of the Web? If so there are huge technical, legal, policy implications” (Carpenter-Negulescu, interview).

At the Dutch National Library (KB) René Voorburg (crawl engineer and coordinator of web archiving) explains that the KB is focusing on storage solutions. Several institutions have coordinated with the KB to hold archives of their web documents, and they have a few in-house generated archives, but none are accessible to general users. Quality assurance, metadata standardization, and standardization of collection criteria are the main focus.

The technical obstacles for individual researchers are similar to those of institutional developers, but of a different scale. Foot took time in her interview to describe several complicated options for researcher workflow needed to create individual archives. Each option solved some problems along the way, but quickly devolved into the technical complications that make most of these desktop solutions less than ideal.

Each individual tool for personal desktop archiving has a different set of goals and so different design elements. Simply archiving sites visited during a particular research session by saving a page's HTML, printing to PDF, or clipping to a tool such as Evernote or Zotero does not always meet the needs of the researcher, and does not scale well beyond small projects. When using desktop tools enable the user to download a copy of a website, often the researcher does not know what metadata elements are missing, or what indexing elements are not accounted for in a particular desktop 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 a mess of archived sites with no logic of how the individual objects can be related to one another. Personal desktop archiving tools are generally designed from a “basic needs” perspective. The designer's assumption is that the user

simply wants to save websites to view later, and does not necessarily consider the social scientists' interest in evidence of a network, connections, and action with regard to web objects.

Neither of these assumptions touches upon the complexity of what a social science researcher thinks it means to save a website or collections of websites for retrospective study. This is reminiscent of Foot's interview description of problems in inter-institutional collaborations in web archiving. People from different disciplines have different concepts in mind even when they use the same terms. These differences can surface in the design of personal desktop archiving tools. It is important to bring those differences to the surface early, and for researchers to be very clear about their research goals, and about what metrics they will need to reach their goals. It is also important for archivists working collaboratively to be clear about their institutional missions, their legal policies, and strategies for collection development. Helen Willa Samuels (2000), originally writing in 1986, claimed, "Our modern, complex, information-rich society requires that archivists reexamine their role as selectors" when it comes to the nature of documentation and records of the past. This need for reexamination given increasingly complex information environments is more relevant than ever, and might also be applied to researchers as they experience the rapidly changing nature of network culture and develop new tools of studying it. Perhaps it is also important to develop some tools that are not multi-purpose - perhaps not all tools need to be accessible to the casual user, and special research tools can be designed to meet the higher level basic needs of the researcher.

"Even with 10 years of experience...it is still a challenge," Foot explained as she described several challenges that she has encountered and watched others struggle with as well. Certain questions cannot be answered, concepts cannot be illustrated, and methods cannot be used if certain metrics, or elements of web objects, are not accounted for as an archive is built; further, studies cannot be replicated if the ephemeral digital primary materials are not 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" (Foot, interview). Julien Masanès of the Internet Memory Foundation echoes these concerns explaining that the Internet Memory Foundation has little feedback from its end-users, who are primarily researchers and cultural "memory" institutions, because researchers often find it "hard to decide in advance what they want" (Masanès, interview).

The second challenge in use is organization. The structure of what is collected matters profoundly. 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, interview). Masanès identifies two areas of tool development that are a primary focus, and are central to making web archives more useful for researchers: 1. Interfaces for navigating and using web archives; and 2. Indexing and search because tools for more detailed analysis require good indexes.

Building tools to index and enable access to such web resources is a complicated and collaborative process. Researchers, archivists, and technologists all highlight the fact that researchers have trouble deciding on what they want methodologically before they begin. This is a recurring missing link in the collaborative effort to build up web archiving tools and practices, and signals that Internet research methods are not primarily simple adaptations of offline methods to online spaces. Certainly there are important questions to be answered with simple adaptations of methods to online spaces, but as Foot reminds us, “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 too late” (Foot, interview). When the researcher collaborating with archivists and technologists discover this conundrum early in the process, there are few resources in the field of Internet research methodology to draw on for guidance to continue the development of research tools.

On Users

In the British Library, Helen Hockx-Yu describes their imagined interaction with users as “We tell them what’s possible and we want them to tell us what’s useful.” How this plays out in practice, however, is less than clear. In identifying his needs as a user of web archives, Enayat described the challenge users face in telling archivists what they want. He described the tool he specified for build in a research project to archive and analyze web resources.

“We wanted it to be open source, and lightweight--in something like PHP. And the work of doing this had to be crowd-sourced, so it was important to build it on something like Drupal that has user management” (Enayat, interview).

He also identified personnel, legal and ethical problems to be addressed in the project build.

“You couldn’t just have one person sitting there all day looking for stuff. And if it is crowdsourced, there has to be some mechanism for identifying and removing duplicates. One of the main problems was the content in Facebook. We were going to anonymize it, of course, but still--do we have permission record something from a friend’s Facebook page?” (Enayat, interview).

Enayat explained his ideal tool for a research-specific web archive.

“My ideal tool would be open source, and written in something like PHP. We would want a reputation system, for example, which may not be a priority for other people, so if it is open we

can build that in ourselves. It could be a web service that you login to and go crawl things, but it should be a mix of automated and manual. You know it is so easy now to script search with Google APIs--give me everything that comes up in the last 24hrs on a particular search term--that it would be great to crawl things and then have something go through them to sort them” (Enayat, interview).

Enayat is an advanced user of web archives, although he requires a team of differently skilled individuals to achieve his research and technical goals.

The Internet Archive has identified three categories to describe their current users. These categories influence their web archiving program: general interest users looking for specific content (e.g., a recipe, blog post, or news story) that is no longer online; personal interest users who are hoping to recreate their own past web presence whether as a result of a loss or for additional back-up; and “Armchair Historians” who do more intense surfing of the Web and can navigate a web archive easily. This collection of projected users does not include scholars or less casual researchers. This is a much broader approach to web archiving than is apparent in other practices building in more traditional institutions.

Wendy Gogel (manager of digital content and projects at Harvard University Library), explains that the user base for Harvard Wax (Web Archive Collection Service) is essentially the curators and collection managers. Harvard Wax is a service offered by the library to Harvard’s Schools and Departments for a fee, and does not include evaluative follow-up of the user. Harvard Wax is an impressive service that shows clearly the state of the art of web archiving. With a two year setup time, collaboration with IIPC, 1.5 FTE in programming effort for two years, plus project management, and a one-time set-up and yearly maintenance fee for Schools and Departments taking advantage of the service, this is a resource-intensive project that offers a fraction of what researchers list as ideal features for web archives. Harvard Wax is an advanced project that clearly shows how far we have come in web archiving, and the cost of how far we still need to go to meet researchers’ needs.

On Solutions

Community building is a particular set of solutions that can pave the way for advancing technology and practices. Helen Hockx-Yu suggests that students funded to address particular questions could lend definition and direction new or currently undefined areas of development. Development of standards could create a common point of departure for new innovative tools, rather than the current paradigm of a common set of tools used in a variety of ways leading to inaccessibility, and little interoperability. These policies, Hockx-Yu argues, can limit the actions of archivists, or even create

barriers to valuable sources for research. Without more exploration into the legal concerns made visible by web archiving, each team in each new institution is left to develop their own policy.

Technical issues are a fundamental, incredibly layered and complex. On the research side, technical solutions are tailored to a specific project. On the libraries and archives side solutions are built broadly to serve a range of general users. Solutions need to come in small bites that can be remixed for specific project needs.

“...there are significant preservation issues for websites on which the digital preservation community has yet to reach agreement. Storage is the least of our problems. The main issue is whether we can provide continued access to the copied websites, and whether we can continue to replicate the functionality and behaviors of our copies; or whether we even need to do that, and if it's sufficient just to preserve the content. Maybe we need to think about preserving versions of browsers” (Pinsent, interview).

When asked what could be considered high priority areas for investment, researchers explained that there needs to be support for infrastructure, for individual archiving, and a way to connect the two. There is a struggle between librarians who identify the needs of a general category of user that is not well-specified, and this does not always involve talking to researchers as users about their specific needs.

“I’d like to see more funding put toward actual facilitating of user communities or at least user consultants in those large infrastructure developments. At the same time it’d be great to have funding directed to individual scholars who are not archiving experts but are willing to help produce archives that could become useful for others to use to help develop that user end of the pipeline” (Foot, interview).

Foot suggested that, “if we were able to equip 100 social researchers across the country, or various countries, with multifaceted toolkits that they could use, that are ideal for the work that they do, there would be some fantastic knowledge gained that could then feed into the development of infrastructure for repositories, as well as develop best practices resources for other social researchers.” Foot suggested a study where the subjects are researchers using web archiving tools where the outcome would be not only sets of collections, but evaluations of tools, and emerging norms around constructed practices for the aggregate collections that she thinks are needed long term. Her work in social science research, that which involves building web archiving and analysis tools, is funded in part by private foundations. In these instances the funders were persuaded that archives would be necessary to generate the kinds of research outcomes expected. “So” she said, “the private funders were not necessarily out to fund archives, they were out to fund other outcomes and other types of research and were persuaded by our group or researchers that archives were necessary in the process” (Foot, interview).

Steven Schneider uses a similar vision when supposing possible solutions. As a researcher, he tackles web archiving as a scholarly method—a valuable and necessary step in the path to other research outcomes.

“A shared platform for social science research [could] give everyone access to a suite of tools to allow end-to-end access to tools: identifier, collector, verifier, coder/annotator, interface (i.e., WebArchivist). Then someone could allow a social scientist to do a small project for \$1000-\$10,000 and collect / analyze / present from 10 - 1,000 sites over a 1 month - 1 year period. We [would] need funding for several years, and a connection with a library that would promise to house / host the collection permanently. That would be cool!” (Schneider, interview).

Discussion

Library and information science fields have been developing practices for collection and archive development for decades, and these practices have come to dominate web archiving. In some ways, the practices and standards of these disciplines are a perfect fit for web archiving. They are extensively developed, standardized, and are ready to handle the content management and delivery systems required across different media types. 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 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).

Library and information science norms have been the basis for many developments in web archiving policy and infrastructure. The result is a strong focus on tools, an archival viewpoint, and traditional modes of collection development relying on broad notions for how the Web archives will eventually be used. This is not unusual; libraries and archives have a broad base of users to accommodate, but in trying to serve the many, there is a risk that few will get the variety of details they really need. For researchers in humanities and social science, this can be the difference between a successful conclusion or an abrupt end to a project. The most basic concerns in traditional modes of collection development may only capture the most basic set of data where a social science or humanities research is concerned. To complicate this

situation further, each additional level of complexity in a web archive may vary given each new research project engaging with the archive. It is no wonder that with all these variables web archivists working to build large-scale collections are focusing on capturing the few variables that are constant across most cases—a quality capture retrievable by either a full-text search or specific URL. No one can find web archives useful if they do not represent those elements, unfortunately, few researchers can find them useful if those variables are the only elements present.

Large libraries and archives continue with their efforts to build large multi-purpose web archives that further institutional missions, building the infrastructure and practices for this endeavor piece-by-piece. Researchers - either on their own, or partnering with archivists - develop their own project-specific archives for use in their research. Archives and other cultural heritage institutions 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 among those with advanced technical expertise, but there is no underlying infrastructure in place to support the research into these archives.

Viewing the Web archive as a collection of documents and bibliographic records is an efficient approach to storing and preserving the Web. Whether the resulting stored data is flexible enough to accommodate the uses to which researchers will want to put web archives is another question. This has set up a point of contention between archivists, librarians and information scientists, and other practitioners 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 collections that will last for a very long time, 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 design new ones. The longevity of the data beyond their own career or even beyond a project, for researchers, is generally of secondary importance to answering the research question at hand.

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.

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.

Conclusions

As outlined above, there are several approaches to building web archives - some developing from institutional mission statements, some from frustration with existing resources, and all from limited understanding of the end-users' needs. Temporary ad-hoc practices that are developed to circumvent obstacles were discussed in several interviews. All respondents described similar obstacles despite their disciplinary background. 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.

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. 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 evidence of 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 along with research methods in humanities and social science.

Stewardship of cultural heritage is a story of loss and reconstruction. Artifacts deteriorate, or become otherwise corrupted, and stewards of the cultural heritage those artifacts represent - whether they be scholars, curators, archivists, or interested amateurs - feel a responsibility to reconstruct not only the artifacts, but often the meaning the artifact 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 artifacts that help researchers to construct those narratives.

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. The restrictions that arise from authoritative management of knowledge can be minimized 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. Encouraging a strong sense of community and shared practice between researchers in humanities, social science, and library and information science can lessen the burdens of overcoming obstacles facing web archiving. Or, as Julien Masanès of the Internet Memory Foundation suggested when interviewed, “what we need is a CERN for web archives.”

References

- 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.
- Bearman, D. & Hedstrom, M. (2000). Reinventing archives for electronic records: Alternative service delivery options. In R. Jimerson (Ed.), *American archival theory: Readings in theory and practice* (549-568). Chicago, IL: The Society of American Archivists.
- Benardou, A., Constantopolous, P., Dallas, C. & Gavrilis, D. (2010). Understanding the information requirements of arts and humanities scholarship. *International Journal of Digital Curation* 5(1). doi:10.2218/ijdc.v5i1.141.
- 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.

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.

Deormedi & Shallcross (2011)

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*. Report. London: JISC. Retrieved from <http://ssrn.com/abstract=1714997> and <http://ie-repository.jisc.ac.uk/544/>.

Dougherty, M. (2007). *Archiving the Web: Collection, documentation, display and shifting knowledge production paradigms* (unpublished doctoral dissertation). University of Washington, Seattle.

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 Revisiting Research Ethics in the Facebook Era: Challenges in Emerging CSCW Research, Savannah, GA.

Dougherty, M. & Schneider, S. M. (2011). 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. (2006). *Web campaigning*. Cambridge, MA: The MIT Press.

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

Gilliland-Swetland, A. (2000). Digital communications: Documentary opportunities not to be missed. In R. Jimerson (Ed.), *American archival theory: Readings in theory and practice* (589-606). Chicago, IL: The Society of American Archivists.

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

Henry, L.J. (2000). Shellenberg in cyberspace. In R. Jimerson (Ed.), *American archival theory: Readings in theory and practice* (569-588). Chicago, IL: The Society of American Archivists.

Kaplan, E. & Mifflin, J. (2000). "Mind and Sight": Visual literacy and the archivist. In R. Jimerson (Ed.), *American archival theory: Readings in theory and practice* (73-100). Chicago, IL: The Society of American Archivists.

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.

- Kluver, R., Jankowski, N., Foot, K., & Schneider, S. (Eds.) (2007). *The Internet and National Elections: A Comparative Study of Web Campaigning*. New York: Routledge.
- Koehler, W. (2004). A longitudinal study of Web pages continued: a consideration of document persistence. *Information Research*, 9(2).
- Masanès, J. (2006). *Web archiving*. Secaucus, NJ: Springer-Verlag New York, Inc.
- Masanès, J. (2013, April 22-26). Leveraging Social Web To Propel Your Archiving Campaign. International Internet Preservation Consortium General Assembly, Ljubljana, Slovenia.
- Mayer-Schönberger, V. (2009). *Delete: The virtue of forgetting in the digital age*. New Jersey: Princeton University Press.
- McCown, F., Marshall, C. C., & Nelson, M. L. (2009). Why web sites are lost (and how they're sometimes found), *Communications of the ACM*, 52 (11).
- McCown, F., & Nelson, M. L. (2009, June 15-19) A Framework for Describing Web Repositories. Paper presented at the Joint Conference on Digital Libraries '09, Austin, TX.
- McCown, F., Smith, J. A., Nelson, M. L., & Bollen, J. (2006). Lazy Preservation: Reconstructing Websites by Crawling the Crawlers, Proceedings of the 8th ACM International Workshop on Web Information and Data Management (WIDM 2006), p. 67-74.
- McEnery, T., & Wilson, A. (2001). *Corpus linguistics*. Edinburgh: Edinburgh University Press.
- Meyer, E. T., Thomas, A., & Schroeder, R. (2011). Web Archives: The Future(s). London: IIPC. Retrieved from <http://ssrn.com/paper=1830025>.
- Mitra, A. (2010). *Alien technology: Coping with modern mysteries*. London: Sage.
- 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.
- Samuels, H.W. (2000). Who controls the past. In R. Jimerson (Ed.), *American archival theory: Readings in theory and practice* (193-210). Chicago. IL: The Society of American Archivists.
- Schneider, S.M. (2006). Scholarly Web archiving knowledge base. Last modified December 2009. <http://www.webarchivist.org/~steve/vks/kbase/vks-swakb1.1.html>.
- 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., & Wouters, P. (2009). Web Archiving as e-Research. In N. Jankowski (Ed.), *e-Research: A transformation in scholarly practice*. New York: Routledge.
- Schneider, S. M., & Foot, K. A. (2010). Object Oriented Web Historiography. In N. Brügger (Ed.), *Web History*. New York: Peter Lang Publishing.
- Shallcross, M. (2011). On the development of the University of Michigan Web Archives: Archival principles and strategies. *Society of American Archivists Campus Case Studies, Case 13*. Retrieved from

<http://files.archivists.org/pubs/CampusCaseStudies/Case13Final.pdf>.

Starbird, K. & Palen, L. (2012). (How) will the revolution be retweeted?: Information diffusion and the 2011 Egyptian uprising. In Proceedings CSCW'12 Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work. Pgs. 7-16. New York: ACM.

Taylor, M. K., & Hudson, D. (2000). "Linkrot" and the usefulness of Web site bibliographies. *Reference & User Services Quarterly*, 39(3), 273-276.

Taylor, J. R. and Every, E. J. v. (2000). *The emergent organization: Communication as its site and surface*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

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.

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>.

¹ <http://www.mementoweb.org>

² <http://fbmya01.barcelonamedia.org:8080/future/>

³ <http://pandora.nla.gov.au>

⁴ <http://lcWeb2.loc.gov/diglib/lcwa/html/lcwa-home.html>

⁵ http://www.kb.nl/hrd/dd/dd_projecten/Webarchivering/index-en.html

⁶ <http://www.sino.uni-heidelberg.de/dachs/>

⁷ <http://handle.library.cornell.edu/VRC/>

⁸ <http://www.bl.uk/aboutus/stratpolprog/digi/webarch/index.html>

⁹ <http://wax.lib.harvard.edu/collections/collection.do?coll=61&lang=eng>

¹⁰ <http://wax.lib.harvard.edu/collections/collection.do?coll=101&lang=eng>

¹¹ <http://pandora.nla.gov.au>

¹² <http://pandora.nla.gov.au/guidelines.html>

¹³ <http://webarchives.cdlib.org/>