Routledge Advances in Research Methods

1. e-Research
Transformation in Scholarly Practice
Edited by Nicholas W. Jankowski

e-Research

Transformation in Scholarly Practice

Edited by
Nicholas W. Jankowski

Routledge
Taylor & Francis Group
New York London

11 Web Archiving as e-Research

Steven M. Schneider, Kirsten A. Foot and Paul Wouters

WEB ARCHIVING AS A FORM OF INQUIRY IN E-RESEARCH

As the Web emerged since the mid-90s as a distinct media form, scholars have increasingly viewed it as an object of study. To facilitate this work, some scholars have turned to Web archiving as a technique and approach, valuing the potential to complete developmental and retrospective analyses of many kinds of online phenomena. Simultaneously, Web archiving has also emerged as a practice of e-research, in which humanities and social science scholars mediate their work via digital and networked technologies. These developments pose challenges for scholars as they seek to develop methodological approaches permitting robust examination of Web phenomena. Some of these challenges stem from the nature of the Web, while others are associated with institutional structures and traditional patterns of behavior of individuals within different types of institutions.

The Web is a distinctive mixture of the ephemeral and the permanent (Schneider & Foot, 2004). There are two aspects to the ephemerality of Web content. First, Web content is ephemeral in its transience—it can be relied upon to last for only a relatively brief time. From the perspective of the user or visitor (and, especially, the researcher), there is little that can be done without specialized tools or techniques to ensure that content can be viewed again at a later time. Second, Web content is ephemeral in its construction—like television, radio, theater and other "performance media," (Hecht, Corman, & Miller-Rassulo, 1993; Stowkowski, 2002) or performance art. Web content once presented needs to be reconstructed or re-presented in order for others to experience it. Web pages are routinely (and increasingly) constructed by computers without human intervention—servers and browsers request, transmit, receive, and process http requests to create the experience of the HTML page—and the activities are repeated (at least in part) when the page is presented again. In other words, the Web is not easily archived in the way that, for example, printed materials are. Books, film, and sound recordings, for example, can be collected in the form in which they are presented; no affirmative steps are needed to re-create the experience of the original; and indeed, when taken such steps cast doubt on the authenticity of the re-presentation.

At the same time, the Web has a sense of permanence that clearly distinguishes it from performance media and performance art. Unlike theater, or live television, or radio, the objects or components that are assembled and then presented as an HTML page—the images, texts, and HTML code—must exist in a stable form prior to their presentation in order to be experienced. The Web shares this characteristic with other forms of media such as film, print, and sound recordings. The permanence of the Web, however, is somewhat fleeting. The traditional Web site regularly and procedurally destroys its predecessors each time it is updated by its producer (indeed, even the blog-style site, which implies an internal and automatic archive, rarely allows a re-creation of the browsing experience at a point in time, opting instead for a re-presentation of content within an alternative experience). In short, absent specific arrangements to the contrary, each previous edition of a Web site is erased as a new version is produced. By analogy, it would be as if each day's newspaper were printed on the same piece of paper, obliterating yesterday's news to produce today's.

The ephemerality of the Web requires that pro-active steps be taken in order to allow a re-creation of Web experiences for future analysis. The permanence of the Web makes this eminently possible. Although saving Web sites is not as easy as, say, saving editions of a magazine, archiving techniques have been and continue to be developed in such a way to facilitate scholarly research of Web-based phenomena. These techniques allow the Web to be preserved in nearly the same form as it was originally 'performed,' similar to recordings of television and radio performances. At the same time, as Web-based phenomena evolve into ever-more dynamic performance, Web archives begin to share the challenges associated with object-oriented representations of other performance media and art (Rush, 1999).

Concerns over the fragility and ephemerality of digital materials have been expressed for almost twenty-five years. For example, an oft-quoted early statement from a 1985 U.S. government report warned that the "the U.S. is in danger of losing its memory" as governments, businesses and other institutions shifted from paper to electronic records (Nelson, 1987); similar concerns were expressed in Europe at about the same time. With the advent of the Web in the early 1990s, this concern grew significantly as greater numbers of institutions and individuals began producing documents in digital form only, rendering traditional modes of archiving less reliable as instruments to preserve records of social phenomena. By 1995, an opportunity to address this concern emerged with the development of Web harvesting programs. Web harvesters or crawlers are applications that traverse the Web following links to pages, initially from a set of pre-defined seed URLs. Harvesters were initially developed by search engines such as Alta Vista to overcome the increasingly impossible task of indexing the Web through human cataloging techniques. Web harvesting technologies gave life to the notion of archiving Web materials via the Web itself.

Three pioneering efforts to archive the Web appear to have developed nearly simultaneously. In 1995, the National Library of Australia, noting that the growing amount of 'born digital' Australian information required its attention, established a working group to select materials for collection and develop techniques for archiving. Out of this initiative came the project whose name, PANDORA, encapsulates its mission: "Preserving and Accessing Networked Documentary Resources of Australia" (National Library of Australia). The PANDORA project archived its first Web site in October 1996, and by June 1997 the Archive contained 31 titles (Cathro, Webb & Whiting, 2001). In early 1996, American entrepreneur Brewster Kahle founded the Internet Archive, a non-profit, private organization whose similar mission is to provide "permanent access for researchers, historians, and scholars to historical collections that exist in digital format" (Internet Archive, 2008; Kahle, 1997). Internet Archive began crawling in 1996, and made its archive publicly accessible in March 1997, when its database contained about two tetrabytes of Web data (Kahle, 1997). In September, 1996, the Royal Library of Sweden launched the "Kulturarw3 Project" to "test methods of collecting, preserving and providing access to Swedish electronic documents" (Mannerheim, 1998). The project completed its first crawl of Sweden's domain in January 1997.

The impetus behind Web archiving activity was clear: The Web was doubling in size every three to six months from 1993 to 1996, and it appeared that the Web had the potential to become a significant platform on which a wide variety of social, political, scientific, and cultural phenomena would play out. Individuals at different types of institutions, such as libraries and archives, whose mission included the preservation of cultural and historical artifacts and materials, recognized the challenge that digital materials presented. Scholars, on the other hand, were a bit slower to recognize the utility of Web archives. Greene et. al. (2004) published one of the earliest efforts to describe Web archiving to support scholarly research. Other early adopters include Hine (2000). Clearly, the Internet Archive anticipated scholarly uses, and the "Energizing the Electronic Electorate" team at the Annenberg Public Policy Center developed tools and techniques in 1999–2000 that formed the intellectual foundation for WebArchivist.org, a research group supporting scholarly Web archiving founded by two of the co-authors of this chapter (Schneider & Foot, 2000; Schneider, Harnett, & Foot, 2001).

Independent of the development of the Web, but strongly related to the development of new Internet applications for science, a new paradigm of research emerged, initially called e-science. This new model of research practice and funding emerged from the communities of supercomputing and physics (Berman, Fox, & Hey, 2003). The core idea was that it should be possible to use the Internet not only as a medium to distribute information, but also as a medium for powerful computational solutions. The metaphor of the 'Grid' exemplifies this by suggesting that one could plug a

particular scientific puzzle into the network and get the solution returned, like one can draw electric power from the power Grid via a socket in the wall. This led to a definition of e-science as the convergence of huge distributed datasets, high-performance computation, and big data pipes for communication, data transfer, and visualization. The first modern Grid was created in the U.S. in the I-Way project in 1995 (Berman, Fox, & Hey, 2003). By 2005, the scope of e-science had spread to the social sciences and humanities and was redefined first as Internet research (Jones, 1999: 1–25), and later as e-research (see Chapters 1 & 3 within). This broader definition of e-research created new connections between scholars involved in the creation of new methodologies in e-science and e-research and the community of researchers who had taken up the Internet as their focused object of study (Brügger, 2005; Schneider & Foot, 2004).

We are only at the beginning of the development of Web archiving as both object and instance of novel research methodologies. Yet, we can delineate the most promising venues for the next few years. Web archiving is in the first instance mainly about Web data: securing access to Web data, long and medium term storage of Web data, and interpretation and annotation of Web data. Data happen to be the single most central concern in e-science and e-research, if we define data in a broad sense—including numerical datasets; complex objects; stored ethnographic observations in the forms of text, audio, and video; traces of individual and collective behavior; and digitized, multi-media historical sources. This confluence will create new possibilities for Web archiving as well as for e-research more generally. As we will see, these will have implications beyond the area of data for research and may redefine defining aspects of public scholarship.

WEB ARCHIVING AND E-RESEARCH PRACTICES

From a social research perspective, the primary reason to archive Web objects is to ensure future access to artifacts that may have analytical value. Web archiving is essential for retrospective analyses of Web objects and developmental analyses of the evolution of Web phenomena over time (Foot & Schneider, 2006). In view of our discussion above of the ephemerality and durability of Web objects, we argue that researchers should archive objects of interest with the assumption that they are potentially dynamic. Assessments of whether, how, and how frequently particular Web objects are actually dynamic may be foundational to other strands of inquiry.

Web archiving is useful for both in-depth studies of a small number of Web objects and broader, large-scale studies as well. For example, scholars interested in the development of one particular site, or of the co-evolution of several sites during a particular timeframe, reap benefits from periodic archiving of their focal site(s) by being able to retrace the emergence of key

features, or changes in certain texts or images over time and in relation to other Web objects and offline phenomena. Scholars who pursue research questions that require large datasets also benefit from Web archives, as they enable many forms of structured data mining and correlational analyses of elements such as texts and links.

Web archiving does not by definition privilege qualitative or quantitative methods of analysis. Whether the data corpus generated through Web archiving processes is relatively small or large, either structured or unstructured forms of observation and analysis may be employed in relation to archived materials. Highly structured observations of particular elements within a Web archive enable broader quantitative analyses; more loosely structured or unstructured observations enable richer qualitative analyses. In contrast to data collection practices that entail extracting HTML text from Web pages, or cataloging links that originate anywhere within a site domain without preserving the structure and texts of the site, archiving collections of pages associated by domain or some other heuristic, and including their hyperlinked context, enables many kinds of data collection and analyses. A well-designed Web archive enables researchers to pursue questions regarding the prevalence of particular elements and the rhetorical function of those same elements within a single study.

Many studies of social phenomena on the Web to date have been based on observations conducted at a single point in time, or during a more extended period but without explicit consideration of the possibility of changes during that period. As mentioned above, Web archiving expands the scope of potential research by enabling developmental analyses of a Web phenomenon as it changes over time by tracking changes in Web objects between rounds of capture in the archiving process. It also expands the scope by supporting cross-sectional research, that is, studies of similar phenomena in different geographical/virtual places, or different periods of time. For instance, a comparative study of the Web productions of several scientific institutes within one country or across different countries during a particular period would be greatly enhanced by robust archiving of all the Web pages within those institutes' domains and their outlinked pages. Similarly, Web archiving affords researchers the ability to compare the Web objects produced in relation to routine events such as a national election with those produced during one or more subsequent events (Jirotka, Procter, Rodden, & Bowker, 2006; Teasley & Wolensky, 2001). These kinds of scope expansions, whether pursued through qualitative or quantitative methods, can increase the depth and strength of the research findings and the overall value of the research.

One common observation regarding e-research is that it may enable new forms of collaboration between geographically distributed researchers (Collins, 1992). Although there are many ways that scholars working alone may benefit from archiving Web artifacts, Web archiving enables and/or enhances some forms of collaboration in the study of Web phenomena that

might not otherwise be possible or efficient. For example, scholars studying the manifestations of a phenomenon in different languages on the Web may find the possibilities for reciprocal translation and joint analysis greatly enhanced by creating a mutually accessible Web archive. As another example, one reliable way to investigate different cultural and/or governmental influences on Web site content would be for collaborating scholars situated in different countries to archive the same Web pages at the same times, utilizing servers from distinct locations, and to then integrate their collections into a single archive for comparative analysis. In addition to these forms of collaboration, various aspects of the processes of Web archiving may be distributed among geographically disparate teams, thus allowing an increased scale of research beyond what any one researcher or team could manage singly.

In sum, Web archiving has been established in recent scholarship as a promising, and sometimes necessary, approach to e-research concerning social phenomena on the Web. Some forms of inquiry are only possible if scholars anticipate the potential evolution of Web objects and establish a prospective archiving regimen. At the same time, the collection of extant artifacts always results in new artifacts, including the collection itself. Therefore, the creation of Web archives is itself a novel form of knowledge production and representation, as we explain in the next section.

WEB ARCHIVING AS FOUNDATION FOR E-RESEARCH

We suggest that scholarly Web archiving proceeds through a set of distinct series of actions or operations on the path from conceptualizing an archive to making archived objects and associated metadata available to researchers or consumers of research. Considering each process discretely provides the opportunity to examine distinct techniques associated with each process, as well as the challenges posed in implementing each process within a scholarly research project.

The first process, *identification*, includes the steps necessary to make known to an archiving system those Web objects to be considered part of the project. For example, a project examining a set of ten Web sites would fully identify the precise definition of the site, as well as the URL of the page from which archiving activity was to begin. Web objects can be identified for inclusion in a project in a variety of ways, including by experts (researchers) or by processing query results from search engines. Specifying an identification protocol that yields a representative and replicable sample of Web content is a considerable challenge. The dynamic nature of some parts of the Web also poses a challenge to the process of identification.

Having identified a set of objects to be included in an archive, the next process, *curation*, involves creating the set of rules and procedures necessary to collect the desired objects, and to verify that the collected objects

match curatorial objectives. These rules might specify, for example, the instructions to be given to the collection process concerning whether to follow or ignore links to other objects, or the disposition of objects inadvertently collected. These rules involve the researcher in ethical dilemmas concerning permissions and notification to producers and distributors of objects to be collected. The process of curation includes the creation of metadata, associated with the identified objects, that is necessary for the collection of those objects, and the verification that collected objects match curatorial objectives.

The process of *collection* encompasses obtaining and storing representations of identified objects. Given a set of identified objects and the rules and procedures to guide the collection, the specific techniques involved with the actual processes of obtaining 'bits on disks' can be determined. The demands of the project and the curatorial objectives will lead an archivist to select from among a variety of collection techniques. For example, an individually-managed collection technique implies that the researcher has assumed the responsibility for the actual collection and storage of archived objects. A system-managed collection relies on software and systems that are administered or managed by individuals or agencies other than the researcher. Consideration of these two collection techniques draw our attention to the locus of control of data collection and suggest potential tradeoffs between robustness and replicability on the one hand, and dynamism, responsiveness, and validity on the other.

Some collection techniques support on-demand collection, suggesting that a researcher can initiate collection of a set of identified objects at any time. On-demand collection has advantages for the research process, as it allows dynamic collection and can be responsive to changes in either the researcher's discovery patterns or to real world events. At the same time, it may short-circuit curation processes (especially object specification) and call into question the systematic nature of the identification processes. One-time collection implies a single crawl or snapshot of identified objects. One-time collection precludes post-crawl or iterative object evaluation or specification. By contrast, periodic collection implies multiple crawls of a set of identified objects at scheduled and defined intervals over a period of time. This approach allows for the comparison over time of archived objects, as well as for refinements of the identification and curation processes over time.

Two processes are associated with generating metadata about collected objects or groups of objects. *Indexing* is the process of generating metadata algorithmically, while *categorization* is the process of generating metadata through observation and analysis. Indexing can involve developing metadata from one of the available sources of information about archived objects, including the object itself, log files from crawling programs, and data developed externally to the archiving project. For example, URLs of crawled objects can be parsed for keywords, to render sites as tree structures,

to identify protocols, or to predict MIME types. Log files from crawling applications might be examined to predict object MIME type, assess object change over time through analysis of checksums, reveal the path through which the object was selected by the crawling program, or indicate the date when the object was first discovered or most recently encountered through the identification and collection process. Objects themselves may be interrogated through indexing techniques to estimate the number of words in HTML objects, generate word concordances within and among objects, determine the number of images and outlinks, and prepare lists of images and outlinked pages.

Categorization, which by definition involves researchers observing and analyzing all or a sample of archived objects, includes analyses of object features and content as well as the annotation of objects. Analyzing features is a categorization technique in which researchers evaluate objects, most commonly HTML pages, for the presence, absence or level of either specific technical characteristics or opportunities for site visitor action. There may be a static set of features analyzed or a dynamic list of analyzed features developed during the course of the categorization process. Metadata generated by this process will most often be represented quantitatively with ordinal or interval level data and exhaustive and mutually exclusive categories. Researchers analyzing content of identified or collected objects often but by no means exclusively focus on HTML pages to generate understanding about a set of objects and to illuminate patterns across multiple objects. The categorization process may be applied to all objects or to a subset of objects. The content may be evaluated using a fixed or static set of measures, or a dynamic list of measures may be developed during the course of the categorization process. Metadata generated by this process may be represented quantitatively with ordinal or interval level data and exhaustive and mutually exclusive categories, or it may be represented in an open-ended structure.

The process of *interpretation* provides metadata about collected objects, derived through the processes of categorization and indexing, to support sense-making activities such as discovery and search, and to facilitate selected representation of collected objects. This process may include the design and implementation of an interface to a Web archive allowing users to select archived objects for examination or analysis. Providing full-text search of both metadata and archived objects is an interpretation technique especially well suited to presenting unstructured data generated through annotation of objects, as well as providing access to archived objects containing text matching submitted queries. Structured queries requiring users to select among various permutations of metadata generated about identified and collected objects provides access to archived objects in a fixed setting; this technique is especially well-suited to presenting categorical data generated through categorization of object features and content, and indexing of object content and characteristics. Another technique that robustly presents similar categorical

data is open-ended N-way queries using a drill-down approach; in this way, users can graphically construct complex Boolean queries using pre-selected categories and terms associated with object metadata.

Finally, collected objects are made available to archive users through the process of *re-presentation*, in which archived objects are retrieved from a collection and re-presented in a Web browser. Considering this step as a process associated with Web archiving is intended to draw attention to fact that rendering archived objects involves affirmative choices and actions that affect the ways in which the rendering is performed. The techniques associated with this process can be distinguished not only by the technical means of re-presentation but by the approach taken to govern access to the objects, distinguish archived objects from the original objects serviced on the Web, and provide contextual metadata associated with the re-presented objects. For example, presentation of archived objects outside of their hypertextual context—in effect, digitization of born-digital objects—is a technique in which captures of full or partial Web objects are created and presented as images. While this mode of re-presentation preserves some of the information value of the object, the underlying source code and links are not available. Similarly, framing objects within a context that emphasizes its archival context draws explicit attention to the fact that the object is not being served by the original site producer in its original context. Either of these techniques can support the presentation of metadata about the object adjacent to the object itself, thus protecting the original integrity of the object while highlighting the descriptive or analytic metadata associated with the object.

Each of these processes necessarily involves multiple challenges for those constructing Web archives in support of e-research. By highlighting these processes individually, we hope to have identified the multiple types of challenges this undertaking will face. As with any research method, Web archiving is best served when the challenges and techniques are identified and acknowledged in the course of creating and presenting the research it supports.

CHALLENGES AND OPPORTUNITIES FOR WEB ARCHIVING IN E-RESEARCH

Social phenomena are so highly complex and variable, it is difficult if not impossible in terms of external validity (much less desirability) to replicate many studies of social phenomena. When Web archives created and/or employed by scholars are made accessible to others, and research processes are well documented, archive-based scholarship is rendered transparent. It becomes possible for others to attempt to replicate sampling and analysis procedures, or—to formulate it a bit more precisely given the problematic nature of 'replication'—to do a secondary analysis (Lecher, 2006). In the research two of us conducted on the use of the Web by U.S. political

campaigns, archival impressions of all of the Web pages referenced in the study we published as a book were provided on a publicly accessible site (http://mitpress.mit.edu/webcampaigning). It is possible that others may observe the same artifacts, with similar analytical techniques, and come to different conclusions. This raises key questions about the importance and role of particular forms of technological and methodological expertise in social research in general and social studies of online phenomena in particular. The possibility of secondary analysis in archive-based studies may also directly or indirectly shape the ways that scholars conduct their research and document their research practices—either to enhance transparency or toward obfuscation—especially as more funding agencies in various countries are mandating the release of social research data along with or shortly after publication of findings.

Another important issue is the relationship between quantitative and qualitative research. Notwithstanding the regular pleas for combining these modes of work, it remains a challenge for many scholars to be able to produce high-quality work in both quantitative and qualitative research traditions. This has partly to do with skills needed (e.g., statistical and formal network analysis on the one hand, close reading and observation on the other hand), but also with basic issues of economy of research. Preparing data collection and fieldwork is already time consuming if one restricts oneself to either quantitative or qualitative research, let alone if one wishes to combine both. Yet, the combination of both can be very rewarding. A weakness of descriptive quantitative research is often that the context of the data is unclear, making the interpretation of the data problematic. Rich qualitative analysis, on the other hand, may suffer from lack of clarity about the extent to which the phenomenon or pattern one finds is relevant outside of the context of the particular case one has studied.

Web archiving does seem to promise novel ways to combine quantitative and qualitative research in one design. First of all, the fact that the datasets can be huge will enable qualitative researchers to check whether a particular phenomenon or pattern they have found in one particular case also seems to be relevant if one looks at a large number of case studies. This could technologically be supported by "pattern matching" software tools (either in the form of Perl or Python scripts, or in NVivo or Atlas.ti type of tools). Second, qualitative data pertaining to a particular case study can be represented as a node in a network, thereby possibly contributing to a better feel for the 'place' of one's case. Third, Web archiving methods enable researchers to collect a variety of quantitative data as harvested metadata. This minimizes the effort on the side of the researcher while still enabling her to couple her own data (whether quantitative or qualitative) to these meta-data. And lastly, quantitative research designs may be enhanced by exploration of concrete instances (e.g., Web pages) of phenomena about which one has quantitative data.

Web archiving employed to support quantitative, qualitative or a blend of these two approaches, can pay off handsomely, in at least two different ways. First, the production of Web archives can create novel datasets to explore research questions in ways that are currently not possible or prohibitively expensive or cumbersome. Second, once these Web archives have been produced, they can be updated, re-analyzed and recombined. This may help to explore additional research questions in the same areas, to open up new areas of research resulting from these recombinations, or to reflexively study the process of research itself. Monitoring the needs of scholars with respect to Web archiving tools and facilities may be part of the latter. The investigation of the U.S. election processes by studying and archiving Web sites of candidates is a good example of substantive implications of Web archiving for social science research (Graubard, 2004). In this study, novel questions about the role of new media in political campaigns could be answered, as well as their implications for the character of political campaigns in the era of the Web. A related study is the Web archive of Dutch political parties maintained at the University of Groningen (Voerman, 2002). In the humanities, the Digital Archive for Chinese Studies (DACHS) focuses on Chinese Web sites in the framework of sinology at the universities of Heidelberg and Leiden (Dougherty, 2007). The latter project is also a good example of the value of sustained small-scale Web archiving. This refers to the main disadvantage of superficial, large-scale Web archives like the Internet Archive or the European Internet Archive for scholarly research. These archives lack the depth and precision in data capture that are usually imperative for scholarly datasets.

A crucial feature of Web archiving as scholarly method is that the annotation of the Web dataset is not delegated to data specialists, but is a joint endeavor in which researchers set the agenda. In its most ambitious form, the results of the analysis of the Web archive become in their turn part of the Web archive. If this is accompanied by metadata describing who uploaded these results, the Web archive can become a crucial tool in a collaboratory or social networking environment by enabling researchers to find each other, comment on each other's results, and set up new projects. This may already add new value, even if the analysis itself would be unaffected by Web archiving as a method. Additionally, it becomes quite feasible to add automatically harvested metadata to the Web archive and include these in, and combine them with, the human produced analysis or interpretation. The analysis itself can of course be supported by specific research software, be it quantitative (e.g., SPSS or UCINET) or qualitative (e.g., NVivo or Atlas.ti). We would like to emphasize again that there are actually no specific methodological or theoretical requirements for the substantive research: both interpretative and formal research designs can be supported by Web archiving, although the technical and organizational requirements would be rather different in each case (for example, the data entry forms need to be adjusted to both the type of questions and the style

of the researcher). It will immediately be clear that the more ambitious types of Web archives are actually also datasets about the researchers that have created and used them. In other words, the Web archives can be 'inverted' and analyzed as a dataset about a particular research community and its methodological and communicative way of life or about the historical development of a particular field of science or scholarship. Given the recent emergence of Web archives this has not yet happened, but it is clearly on the horizon.

Lastly, a specific case that bridges the substantive and the reflexive dimensions mentioned above is the opening up of scholarship to the public by enabling non-academic communities to annotate Web archives. According to Dougherty (2007), this is the key implication of Web archiving for scholarly research. Her contention is based on the massive popularity of mundane annotation practices in environments such as email lists, blogs, social tagging, and social network sites. By giving comments and feedback to each other as part of their social life, people nowadays are actually annotating their environment on the basis of their own experiences and expertise. One can see this as additional data that can be incorporated in e-research and Web studies. One can also, however, see non-academic audiences as alternative producers of socially relevant knowledge that sometimes can compete in terms of rigor and increasingly also in terms of available resources, with the most serious scholarship (Hackett 2008). In this sense, Web archiving as method in e-research could then contribute to a crucial turn in the demarcation of scholarly expertise and help develop totally new forms of public scholarship (Massanès, 2006). This supports Dougherty's conclusion that a reevaluation of the notion of expertise must take place before roles for different knowledge brokers in digital culture heritage can be identified (Dougherty, 2007). And last but not least, this also leads to a redefinition of what Web archiving really is about. It is not so much a matter of preserving more or less complete datasets or digital experiences, supervised by designated experts, but rather a strategy for creating multiple paths to artifacts and information that is evidenced by nonhierarchical search tools on the Web (Massanès, 2006), thereby enabling multiple ways of analyzing, understanding and making sense of our world.

TAKING WEB ARCHIVING SERIOUSLY: IMPLICATIONS FOR E-RESEARCH

Taking Web archiving seriously as both a novel research methodology and a data collection suite of tools also means that researchers using Web archiving as (one of) their methodologies will have to be supported with sufficient information expertise and human power. As a consequence, either scholarship in the humanities and social sciences will itself have to become more capital intensive, or scholars will need to have secure access

to standardized and personalize-able Web archiving tools in a stable infrastructure, including storage and annotation facilities, provided by the universities or national libraries. We have already indicated that there are good reasons for the development of a combination of two different types of Web archiving: broad and superficial, such as the Internet Archive or the Swedish cultural heritage archive, versus narrow and deep, such as the Digital Archive for Chinese Studies archive. Since we cannot predict which research questions will become pertinent, even in the near future, the production of new Web archives to address particular research questions will keep emerging. This cannot be covered by broad sweeping routine Web archiving, because it is impossible to archive 'the whole Web'. Every Web archive will by necessity always represent a limited sample out of the universe of Web objects. Actually, we do not really know how useful broad, ill-defined Internet archives will be. We do know, however, that social scientists and humanities scholars will for many research questions need focused and well-defined Web archives if they need Web archives at all. Therefore, we would like to offer as our intuition that in the next few years it is most important that scholars across a variety of fields be provided with and/or develop the tools and data infrastructures that enable them to create their own Web archives and to share their experiences in this novel type of research with each other. The development of technical and scholarly standards should go hand in hand with this process. We expect more results from this approach than from the creation of large Web archiving infrastructures that are not driven by research questions and needs.

Complementarily, we urge scholars to initiate Web archiving efforts in collaboration with libraries and archives, or to seek to participate in them. Individual scholars can contribute robustly to institutional Web archives by offering their perspective as would-be or actual users or beta testers of proposed or prototype Web archives. Two of us have had fascinating experiences as the only non-librarians at Web archive planning sessions organized by individual libraries or associations of libraries. There can be significant gaps between what librarians or professional archivists think that scholars-as-archive-users want, and what scholars actually want—dialogue between scholars and librarians/archivists about Web archiving is foundational for redressing these gaps.

We suggest that during the design phase of a research project that will entail Web archiving, investigators identify a library or institutional archive that would be willing to collaborate in the archiving process, or at least be a repository for the Web collection during or at the close of the research project. Terms for long-term preservation and access for other scholars and the public should be negotiated upfront. Scholars may need to supply a collaborating library/archive with forms of metadata beyond those required for the research, but we contend that the long-term benefits of durable, accessible Web archives are worth the effort of coordinating. Through such collaborations, scholars may have opportunities to shape the collection

practices, architecture designs, and policy decisions that will have a significant impact on the nature of e-research projects that can be supported, both presently and in the future.

Although Internet researchers still use mainly 'live' Web data that exist in one particular point in time, e-research infrastructures promise to be relevant research environments for scholars using Web archives. This would require the introduction into e-research infrastructures of a flexible Web archiving support structure with considerable scope for the individual scholar and small research group. Even small Web archiving projects tend to produce massive datasets, which can most efficiently be hosted in large data Grids.

So what institutional policies are needed to support and sustain Web archiving as a viable e-research practice? To answer this question, we need to review the current institutional environment of Web archiving, as well as the scope of the implications of Web archiving for social science and humanities scholarship. Most Web archiving projects have been defined in the context of national libraries, pioneered by the libraries of Australia, Sweden, and the U.S. Library of Congress. The Dutch National Library (KB) has recently decided to add a limited experiment in Web archiving to its e-Depot, a pioneering digital facility to preserve digital scholarly publications (Hoorens, Rothenberg, Van Oranje, Van der Mandele, & Levitt, 2007). National libraries tend to focus on the preservation of (digital) cultural heritage and/or the public record of science and scholarship. They are much less oriented towards problems of access to, and sharing of, research data (Arzberger et al., 2004). This means that they also tend to underestimate the complexities of providing researchers with the capabilities to annotate Web archives and to add these annotations to the Web archive. Related to this, as Dougherty (2007) observes in her PhD thesis on Web archiving, "the steps of categorization, interpretation, and representation are overlooked entirely." (Dougherty 2007, p. 42)

The upshot of this is that if national libraries and archives wish to support Web archiving as a research activity, they will need to make a double move. First, the traditional paradigm of archiving, fundamentally based on stable documents, needs to be transformed into a paradigm that enables archiving the permanent/fleeting phenomenon of, as well as phenomena on, the Web. Second, they need to "open up" key elements of the archiving procedures to enable a "natural flow" of research results into the Web archives, both as primary data and as metadata. This requires a rethinking of the responsibilities of archivists and librarians in relation to scholarly research, and vice-versa the integration of key elements of the scholarly production cycle (Borgman & Furner, 2002) into the production of Web archives and their maintenance.

Scientific institutions also need to change, in order to enable researchers to create Web archives as part of their research task and receive the credits for this pioneering work. This is comparable to the need in other fields, like bio-informatics, to recognize the creation of scientific databases as part

and parcel of scientific research. This will have to lead to crucial changes in human resource management and personnel policy at universities and research institutes. Presently, the creation of datasets is merely recognized within a particular research project, but not as a research oriented infrastructural task. It will become imperative that funding agencies recognize these new scholarly challenges as critical for new endeavors in the social sciences and humanities.

CONCLUSIONS

Web archiving is a valuable and critical research method for scholars engaged in e-research. As a method, it will likely become increasingly important and relevant to scholars whose research projects rely on data collected from Web resources. As the breadth and depth of such research expands and as analyses of Web-based objects become both more common and more accepted, scholars will find it desirable to develop techniques to capture their datasets in ways that facilitate replication and enhance validation. In addition, external reviewers and funding agencies will expect such research to be conducted using these techniques.

In order to ensure the reliability and validity of Web archives, careful attention to all of the processes and systems involved is required. As in research involving experiments, surveys, content analysis, or any quantitative or qualitative methods, close attention to and documentation of all processes, choices, and decisions is critical to successful social research. We should insist that Web archives deployed in e-research projects embed rich description of these processes. In this way, future users of archives can address the possible impact of choices made in constructing archives on the conclusions drawn by researchers.

As e-research becomes institutionalized around practices supporting digital repositories and data curation, Web archiving will become one of several research methods supported and encouraged. As such, Web archiving presents unique opportunities for scholars to collaborate with institutions such as libraries and archives, and will impose unique challenges on these institutions to collaborate with scholars. At the same time, individual scholars with fewer resources can successfully employ the method. In all cases, the clarity of the research method is more important than the scale of the research effort.

ACKNOWLEDGEMENTS

The authors would like to thank Charles van den Heuvel, Ernst Thoutenhoofd, Sally Wyatt, and the referees for the comments on an earlier draft of this chapter.

REFERENCES

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., & Laaksonen, L. (2004). Science and government: An international framework to promote access to data. *Science, 303*(5665), 1777–1778.

Berman, F., Fox, G., & Hey, T. (2003). The Grid: Past, present, future. In F. Berman, G. Fox & T. Hey (Eds.), *Grid Computing. Making the Global Infrastructure a Reality* (pp. 9–50). Chichester, West-Sussex, UK: John Wiley & Sons.

Borgman, C., & Furner, J. (2002). Scholarly communication and bibliometrics. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (Vol. 33, pp. 3–72). Medford, NJ: Information Today Inc.

Brügger, N. (2005). Archiving Web sites: General considerations and strategies. Aarhus, Denmark: The Centre for Internet Research.

Cathro, W., Webb, C. and Whiting, J. (2001). Archiving the Web: The PANDORA Archive at the National Library of Australia. 2001. Paper presented at the Preserving the Present for the Future Web Archiving Conference, Copenhagen (June).

Collins, H. (1992). Changing Order: Replication and induction in scientific practice. Chicago: University of Chicago Press.

Dougherty, M. (2007). Archiving the Web: Collection, documentation, display and shifting knowledge production paradigms. Unpublished PhD thesis. University of Washington, Seattle.

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

Graubard, S. (2004). *Public Scholarship: A New Perspective for the 21st Century* (Report). New York: Carnegie Corporation of New York.

Greene, M. A., Boles, F., Bruemmer, B., & Daniels-Howell, T. J. (2004). The Archivist's New Clothes; or, the Naked Truth about Evidence, Transactions, and Recordness. Ann Arbor: University of Michigan Sawyer Seminar.

Hackett, E. J. (2008). Politics and Publics. In Hackett, E. J., Amsterdamska O., Lynch, M., & Wajcman, J., *The Handbook of Science and Technology Studies Third Edition*, MIT Press, Cambridge, p. 429–432.

Hecht, M. L., Corman, S. R., & Miller-Rassulo, M. (1993). An evaluation of the drug resistance project: A comparison of film versus live performance media. *Health Communication, 5*(2), 75–88.

Hine, C. (2000). *Virtual ethnography.* Thousand Oaks, CA: Sage.

Hoorens, S., Rothenberg, J., Van Oranje, C., Van der Mandele, M., & Levitt, R. (2007). Addressing the uncertain future of preserving the past. Towards a robust strategy for digital archiving and preservation. Arlington, VA: RAND Europe.

Internet Archive. (2008). *About the Internet Archive.* Retrieved May 8, 2008, from http://www.archive.org/about/about.php

Jirotka, M., Procter, R., Rodden, T., & Bowker, G. C. (2006). Special Issue: Collaboration in e-research. *Computer Supported Cooperative Work: The Journal of Collaborative Computing, 15*(4), 251–255.

Jones, S. (Ed.). (1999). *Doing Internet Research: Critical Issues and Methods for Examining the Net.* Thousand Oaks, CA: Sage.

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

Lecher, H. E. (2006). Small scale academic Web archiving. In J. Masanès (Ed.), *Web Archiving*, pp. 213–226. New York: Springer.

Mannerheim, J. (1998). Problems and Opportunities in Web Archiving, Nordic Conference on Preservation and Access: National Libraries and Research Libraries in a time of change. Stockholm.

Massanès, J. (Ed.). (2006). *Web Archiving.* New York: Springer.

National Library of Australia. *PANDORA: History and Achievements.* Retrieved May 8, 2008, from http://pandora.nla.gov.au/historyachievements.html

Nelson, A. K. (1987). The 1985 Report on the Committee on the Records of Government: An Assessment. *Government Information Quarterly, 4*(2), 143–150.

Rush, M. (1999). A noisy silence. *PAJ: A Journal of Perfomance and Art, 21*(1), 1–10.

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. (Eds.). (2000). *Annenberg 2000 Election Web Archive.* Philadelphia: Annenberg School of Communication, University of Pennsylvania.

Schneider, S. M., Harnett, B. H., & Foot, K. A. (2001, May 23–28). *Catch and code: A method for mapping and analyzing complex web spheres.* Paper presented at the International Communication Association, Washington, DC.

Stowkowski, P. A. (2002). Languages of place and discourses of power: Constructing new senses of place. *Journal of Leisure Research, 34*(4), 368–382.

Teasley, S., & Wolensky, S. (2001). Scientific collaborations at a distance. *Science, 292*, 2254–2255.

Voerman, G. (2002). Archiving the Web: Political party Web sites in the Netherlands. *European Political Science, 2*(1), 68–75.