Observing the Web by Understanding the Past: Archival Internet Research

Matthew S. Weber
Rutgers University
4 Huntington St. #206
New Brunswick, NJ 07030
848-932-8718
matthew.weber@rutgers.edu

ABSTRACT

This paper discusses the challenges and opportunities for using archival Internet data in order to observe a host of social science phenomena. Specifically, this paper introduces HistoryTracker, a new tool for accessing and extracting archived data from the Internet Archive, the largest repository of archived Web data in existence. The HistoryTracker tool serves to create a Web observatory that allows scholars to study the history of the Web. HistoryTracker takes advantages of Hadoop processing capacity, and allows researchers to extract large swaths of archived data into a link list format that can be easily transferred to a number of other analytical tools. A brief illustration of the use of HistoryTracker is presented demonstrating the use of the tool. Finally, a number of continuing research challenges are discussed, and future research opportunities are outlined.

Categories and Subject Descriptors

J.4 [Computer Applications]: Social and Behavioral Sciences---communication: H.3.4 [Information Storage and Retrieval]: Systems and Software---Information networks

General Terms

Documentation, Design, Verification.

Keywords

Archived data, web observatory, network analysis, Occupy Wall Street, social sciences, data extraction.

1. INTRODUCTION

Across disciplines, archival Internet data represent a vast repository of untapped research potential. For public audiences, the Internet Archive repository has proved immensely popular; the public Wayback Machine (www.archive.org) interface to the Internet Archive serves more than 300,000 visitors a day, and more than 200 requests a second. As of 2013, the Internet Archive contains more than seven petabytes of data and offers a reliable historical record of Web sites dating from 1995 to the present. In terms of data availability, the Internet Archive is by far the largest digital source for historical research pertaining to the Web and its contents over time.

Although the Internet Archive contains billions of Web pages and has tremendous potential to facilitate research, research languishes because the search, crawl and extraction functions are severely limited. For example, to browse Internet Archive resources, one generally needs to know the exact uniform resource locator (URL) of interest. A researcher cannot run a full-text search of the entire history of the Web, nor are there facilities to download sets of related Web pages except for manual download. Despite the potential for research using data from the Internet Archive, there has never been an article published using Internet Archive data in major academic journals such as *Science*, *Nature*, *Journal of Communication, Academy of Management Journal,* or the many other leading disciplinary journals. The lack of publications reflects the current barriers to accessing large-scale data from the Internet Archive.

Thus, this paper introduces the HistoryTracker tool, a resource that allows scholars to extract research ready data sets from Internet Archive content. A primary objective of the HistoryTracker tool is to tear down those barriers and simultaneously bring together a new community of researchers. In recent years, Internet Archive has initiated a series of funded projects aimed at making its data more amenable to large-scale research; this effort has included layering metadata into Internet Archive files. In combination with recent computing advances, such as the development of Hadoop computing technology, an opportunity presently exists to make significant advances toward fulfilling the potential of Internet Archive for social science research.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579213

2. THE POTENTIAL OF DIGITAL ARCHIVE RESEARCH

Digital libraries such as the Internet Archive provide a unique opportunity to trace social phenomena in the online space. In recent years, researchers in a number of disciplines have capitalized on the promise of digital archive data to produce cutting edge research. For instance, political scientists used archived digital content to map the connections between political blogs, illustrating the divided nature of politics online [1]. More recently, subsequent studies have illustrated the divided nature of the blogosphere in a global context using archives of Korean blogs [8] as well as Arabic blogs [11]. Likewise, research using archived Web documents has demonstrated the changing nature of media and political discourse in the online environment. An examination of archived Drudge Report Web pages demonstrated that digital political media were highly dependent on existing traditional media for content [18]. A separate analysis of government documents showed that key stakeholders edited digital records to alter the recorded narrative [19].

In media studies, scholars have also used archived Web data to examine the content of Web pages and to demonstrate the increasing brevity of news content on Websites [12]. In the context of organizations and information, Weber and Monge [31; 32] used sub-samples of data from Internet Archive to examine mechanisms that drive link formation between news organizations, and demonstrated that a few central actors control the flow of news between wire services, blogs, and newspapers. A second study used archival data to demonstrate the use of hyperlinks as a tool for establishing partnerships between organizations [30]. Looking within academic organizations, Thelwall and colleagues [22; 27; 28] have conducted a number of studies using archived Web data to demonstrate the interdependence of academic institutions on the Web, and to illustrate how different academic fields have grown in relation to one another. Similarly, research has shown that an accurate archiving reference system is important for preserving references in scholarly publications [13; 16]. The issue of persistence of references is particularly problematic in Law, where online citations are increasingly used in law reviews [24], and in health, where changing content often translates to differing perceptions of the appropriateness of medical treatment [29].

From the perspective of computer science and related fields, researchers have sought to address a number of questions pertaining to archival Internet data. First and foremost, work in computer science has emphasized the need to develop a rigorous framework for storing digital data. This includes work on indexing, emphasizing the challenges associated with developing a robust indexing method for archived data to accurately track versions [14; 16]. As the amount of content online continues to expand, computer scientists are also working on using existing archives to test and develop improved methods for visualizing and mapping vast amounts of archived data [4]. Better indexing systems and a more accurate understanding of the type of content archived also have the potential to lead to improvements in automatic classification systems [10]. A number of studies have looked at the validity of archives. Researchers examined how much of the World Wide Web is actively being archived and demonstrated that, depending on the type of content, anywhere from 35% to 90% of the Web is being retained in digital archives [2]. Another study demonstrated the global distribution of archived Web pages and the general bias toward Web pages in developed countries [25].

Focusing specifically on the Internet Archive, a number of studies have established the validity of the use of Internet Archive data – and associated tools – as a basis for research in the social sciences. In particular, scholars have used Internet Archive as a tool for estimating the age of a Web site and the frequency of updates, and for evaluating and coding the content within sites [6; 26; 29]. Further, Murphy, Hashim and O'Connor [21] validated measures of age and frequency of updating against third-party data, illustrating the overall strength and reliability of these measures as research tools. Thus, research supports the use of data from Internet Archive as attributes and characteristics in research studies. Internet Archive can also be used as an evolutionary research tool to track the development of technology and track changes in content over time. Chu, Leung, Van Hui and Cheung [9] conducted a longitudinal study of e-commerce Web sites using the Wayback Machine to track the development of site content. Similarly, Hackett and Parmanto [13] used the Wayback Machine to analyze changes in Web site design in response to technological advances over time.

All of this work establishes a strong precedent for research using Internet Archive data and a clear, growing interest in research examining issues that use data from digital archives more generally. To date, however, researchers have used tools that have been time-intensive to develop and that are custom-made for particular topics. As a result, previous tools and research methods are not widely usable.

Large-scale studies using Internet Archive data, and large-scale datasets in general, are hampered by the size of the database, the structure of the data itself, and the complexity of linkages between sites [7; 21]. Regarding the scope of archival data, a number of specific challenges exist. First, the prospect of mining data from a seven-petabyte database creates an imposing challenge. Few tools exist that enable scholars to find the right information and to translate information into a researchable format. Second, researchers face challenges associated with analyzing longitudinal datasets that represent the growth and evolution of content online. Previous research projects seeking to draw on historical perspectives of the Web have found that exact data are hard to ascertain, even when the area of study is so focused that each Web site owner can be individually contacted [5; 15]. One project has achieved an analysis of the dynamics of moderately large areas of the Web (UK university Web sites) but only at the expense of a long-term large-scale data collection exercise, and it delivered only annual data [22].

As scholars from the social sciences engage with large-scale data research, existing research methods are pushed to their limits and new methods are needed – often bridging between fields [17]. Indeed, the computational social science research agenda outlines the clear need for the blending of theoretically driven research with rigorous research methods. Another key challenge in working with data from Internet Archive is the structure of the data itself. Internet Archive is structured implicitly, as patterns of associations and interactions are inferred from the nature of content with the sites themselves rather than by formal structures. This is in stark contrast to traditional relational databases, where interactions are explicitly delineated in specified fields and relationships [3]. Thus, extraction of data from the archives relies on the structure of the archive rather than the structure of selected domains.

3. HISTORYTRACKER

The previous section outlined the major challenges associated with conducting research utilizing archived Internet data. In the following section, the HistoryTracker tool is introduced as a tool for accessing archived Web data in a research ready format. The primary function of HistoryTracker is to allow researchers to extract hyperlink trace data of the connections between web pages.

3.1 Data Format

Historic data before 2009 contained in the Internet Archive is stored primarily in the ARC (ARChive file) format, which was defined by IA in the late 1990s to store the actual content of Web pages, including hypertext markup language (HTML), image files, and graphics. However, beginning in 2009 most new data entering the collections were transitioned to being written in WARC format. WARC (Web ARChive file format) is an ISO standard file format that expands on ARC and adds enhanced metadata records describing the Web content. With the advent of WARC, the metadata was stored in the WARC files alongside the Web content for easy retrieval by access and analysis tools. The use of WARC files streamlines processing of the Internet Archive, and will aide in accelerating research and access.

In addition to the WARC files for each domain, IA is in the process of transitioning to a dual file structure; the entire archive is being reprocessed to create a WAT (Web Archive Transformation) metadata file and WARC archive file for each domain. The WAT file formats the metadata using JavaScript Object Notation (JSON) language, which allows the data to be formatted in an easy-to-parse hierarchy. The WAT file is intended to facilitate efficient large-scale data extraction and is easily integrated with the Hadoop software framework. The metadata for each domain contain a number of key descriptive elements, including:

- general information about the domain, including dates when the domain was archived by Internet Archive;
- outlinks from the domain that were present each time the domain was archived;
- the size of the domain (measuring the amount of content present); and
- a uniform resource identifier allowing for rapid identification of the actual WARC file containing the archived content.

3.2 Tool Functions

HistoryTracker is a data extraction tool designed to access Internet Archive data and extract content and links starting with a seed set of URLs and expanding outward to collect a full dataset focused on a given topic. This tool uses key URLs and keywords selected by the researcher as a starting point and extracts all related domains to create topically based databases for research. For example, take the work of tracing the evolution of www.rutgers.edu over time. A query for http://www.rutgers.edu will extract all known records of the domain and trace back all known links to and from that page with date stamps per link indicating when in time that link existed. In addition, the output includes summary information about the domain, including size and type of content, number of links, and a summary of the top level domains in the extracted data set.

HistoryTracker takes as input a seed list of URLs, from which the HistoryTracker tool will start searching the archives. The first component of the tool is the extractor; this component queries the archived data and extracts the raw links content. The second component is a series of filters that removes spam and formats the data for analysis; this portion of the tool moves the data from a Hadoop file system format to a text based format that is easier for researchers to work with. Each component is detailed as follows:

1. **Extractor:** The primary component of the HistoryTracker tool is the extractor interface. The extractor takes a seed list of URLs as input. The extractor can then search up to three steps out from the initial URL seed data. Researchers are able to designate specific search parameters. For instance, the tool can take the input URLs as a starting point, and then crawl all related links outbound from that list. Alternatively, the tool can also perform a matching task, returning only those links where both the original seed URL and the destination (linked to) URL are contained in the seed list. The second approach is

useful when looking to examine a particular community of websites.

Internet Archive crawls progressively; only a small portion of Web pages is crawled on a given day. If a given Web site such as www.wlv.ac.uk was crawled by IA on March 26, 1997, it may not appear again in the archive until December 1, 1997. Thus the extractor tool examines data in windows, pulling snapshots of the data in 1-month ranges. Researchers are able to narrow this window to look at data from specific dates in the post-extraction phase. A researcher working on a particular topic or community (political movements, research fields, topical news issues, materials preserving cultural heritage and the like) thus begins by defining a collection of URLs, or a set of keywords, that encapsulates the core research interest.

The extractor tool aggregates all links, including spam and sites that are likely to be unrelated to the search topic, and records the associated descriptive data. The data that is returned includes the following: originating (seed) url, linked to url, date the link occurred on, size of the originating web page in bytes, and the number of times the link occurred between the two pages on a given date. Subsequent releases will include descriptive text and keywords.

2. **Filters:** The second component consists of a series of filters that transform the extracted data for research and analysis. The filters are primarily based on open source blacklists that enable the automatic deletion of unwanted advertising, pornography and spam. Because much of the data is historic in nature, it is challenging to accurately filter out unwanted pages. The initial version uses the blacklists as a starting point for filtering.

   This final step also reformats the data for research; data is extracted into a Hive database. From there, the data is extracted a second time into a raw tab separated file format that includes the above listed fields. The researcher is able to specify the exact fields of interest, and customize the exact nature of what is outputted.

The extractor and filter are combined into the HistoryTracker tool. The tool itself is based on Hadoop and Java The incorporation of the Hadoop file structure and servers represents a significant step forward in researchers' general ability to access and manage these types of datasets. In addition, these tools leverage open-source programming and will lead to the development of scalable and customizable tools for data-intensive social science research.

3.3 Community Building and Data Access

In order to facilitate the initial development of the HistoryTracker tool, and to manage access to the data, the initial release of the HistoryTracker tool builds on a number of topical areas for researchers. Initial data sets focus on content surrounding the Occupy Wall Street movement, Superstorm Sandy, the United State House and Senate websites within the .gov domain, and news media websites. Content areas were chosen based on existing research interests, and will be expanded over time. The initial output data and the source code are both available through a community research portal that can be accessed at https://archivehub.rutgers.edu. The web portal provides a rich, community-building environment that features tools for sharing prototypes, hosting and disseminating databases, collaborating, and sharing ideas via blogs and wikis.

Furthermore, this space allows researchers to engage in an ongoing discussion around the many challenges that remain with regards to research on Internet Archive data. The lead research team is using this as a space to maintain transparency about the project, and to gain input as the research shifts to focus challenges such as conducting social network analysis on a large scale and managing large scale semantic data extraction.

4. ANALYZING HISTORYTRACKER DATA

Research advocating large-scale data in computational social science has promoted the use of network analysis as a method for analyzing large sets of interconnected data [17]. Indeed, a number of scholars have suggested that social network analysis is a promising tool for the study of archived Internet data, as it allows scholars to decode connections between websites in a meaningful fashion [7]. Network methods such as exponential random graph (ERG) models provide a means for testing changes in communities of data; ERG models are particularly well suited for testing patterns of connection and interaction within large datasets [23]. In addition, researchers will be able to utilize numerous complementary approaches, including aspects of online ethnography [20] and textual analysis.

The tab separated file format produced by the HistoryTracker tool is particularly well suited for social network analysis. Working with the R framework, and similar interfaces, the tab separated formatted can be imported as an edge list format that is easily analyzed from a social networks perspective. Packages such as NetworkX, iGraph and Statnet each allow researchers to use edgelist data as input, and to quickly analyze the resulting data.

5. FUTURE RESEARCH

The preceding sections have detailed the major features of HistoryTracker, a new tool for conducting archival Internet research. In addition, this paper outlines the potential for archival Internet research to open up new avenues of research.

As a tool for observing the web, HistoryTracker provides a unique mechanism for examining the past of the Internet. There are few resources available today that allow scholars to examine a large corpus of websites over a sustained period of time; HistoryTracker accomplishes exactly that. At the same time, the HistoryTracker tool, and similar research tools being developed by others, raises a host of future challenges.

First and foremost, as scholars increasingly begin to work with large-scale data, new analytical methods are needed to work with the data. The social network analysis packages described above are designed to work with larger data sets, but small extractions from the HistoryTracker tool generate hundreds of thousands of records.

Second, HistoryTracker is an initial tool for data extraction, and it is limited in terms of scope. The tool is designed primarily to provide access to link list data and related attributes. This leaves a host of web content untapped. Researchers are currently working to expand the HistoryTracker tool to include summaries of the text included in the archive files. Challenges remain in processed text summaries for millions of web pages, but there is significant promise in being able to examine decades of language used in digital spaces.

Indeed, significant work is needed to provide access to the semantic content contained within the Internet Archive, and the are many challenges that remain in granting full access to this data. The HistoryTracker project provides a step forward in developing a new research agenda in this space, and creates a unique web observatory for looking at the development of the World Wide Web over time.

6. SECTIONS

The author acknowledges support from the National Science Foundation (NSF Award #1244727), as well as the support of a number of collaborators including Kris Carpenter, David Lazer, Katherine Ognyanova, Vinay Goel, Luan Nguyen, Hai Nguyen and Allie Kosterich.

7. REFERENCES

[1] ADAMIC, L. and GLANCE, N., 2005. The political blogosphere and the 2004 U.S. election: divided they blog. *The proceedings of the 3rd International Workshop on Link Discovery*, 36-43. DOI= http://dx.doi.org/10.1145/1134271.1134277.

[2] AINSWORTH, S., ALSUM, A., SALAHELDEEN, H., WEIGLE, M.C., and NELSON, M.L., 2011. How much of the web is archived? In *JCDL 2011* ACM Press, Ottawa, Ontario, Canada, 133-136.

[3] ARMS, W., AYA, S., DMITRIEV, P., KOT, B., MITCHELL, R., and WALLE, L., 2006. A Research Library Based on the Historical Collections of the Internet Archive. *D-Lib Magazine 12*, 2.

[4] AVCULAR, Y. and SUEL, T., 2011. Scalable Manipulation of Archival Web Graphs. In *Workshop on Large-Scale and Distributed Systems for Information Retrieval*, Glasgow, UK.

[5] BJÖRNEBORN, L., 2004. Small-World Link Structures across an Academic Web Space - a Library and Information Science Approach Royal School of Library and Information Science, Copenhagen.

[6] 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.

[7] BRÜGGER, N., 2012. Historical network analysis of the web. *Social Science Computer Review 31*, 3, 306-321. DOI= http://dx.doi.org/10.1177/0894439312454267.

[8] CHANG, W.-Y. and PARK, H.W., 2012. The network structure of the Korean blogosphere. *Journal of Computer-Mediated Communication 17*, 216-230. DOI= http://dx.doi.org/10.1111/j.1083-6101.2011.01567.x.

[9] CHU, S.-C., LEUNG, L.C., HUI, Y .V ., and CHEUNG, W., 2007. Evolution of e-commerce Web sites: A conceptual framework and a longitudinal study. *Information and Management 44*, 2, 154-164.

[10] DAI, N., DAVISON, B.D., and QI, X., 2009. Looking into the Past to Better Classify Web Spam. In *Fifth International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)* ACM Press, Madrid, Spain, 1-8.

[11] ETLING, B., KELLY, J., FARIS, R., and PALFREY, J., 2010. Mapping the Arabic blogosphere: politics and dissention online. *New Media & Society 12*, 8, 1225-1243. DOI= http://dx.doi.org/10.1177/1461444810385096.

[12] GREER, J.D. and MENSING, D., 2006. The Evolution of Online Newspapers: A Longitudinal Content Analysis, 1997-2003. In *Internet Newspapers: The Making of a Mainstream Medium*, X. LI Ed. Lawrence Erlbaum Associates, Mahwah, NJ, 13-32.

[13] HACKETT, S. and PARMANTO, B., 2005. A longitudinal evaluation of accessibility: Higher education web sites. *Internet Research 15*, 3, 281-294.

[14] HE, J., YAN, H., and SUEL, T., 2009. Compact Full-Text Indexing of Versioned Document Collections. In *18th ACM Conference on Information Knowledge Management (CIKM)* ACM Press, Hong Kong, HK, 415-424.

[15] HOLMBERG, K., 2009. A webometric analysis of local government websites in Finland Proper Åbo Akademi University Turku, Finland.

[16] LAWRENCE, S., PENNOCK, D.M., FLAKE, G.W., KORVETZ, R., COETZEE, F.M., GLOVER, E., NIELSEN, F.A., KRUGER, A., and GILES, C.L., 2001. Persistence of Web References in Scientific Research. *Computing Practices 34*, 2, 26-31.

[17] LAZER, D., PENTLAND, A., ADAMIC, L.A., ARAL, S., BARABASI, A.-L., BREWER, D., CHRISTAKIS, N., CONTRACTOR, N.S., FOWLER, J., GUTMANN, M., JEBARA, T., KING, G., MACY, M., ROY, D., and VAN ALSTYNE, M., 2009. Computational Social Science. *Science 323*, 721-723.

[18] LEETARU, K., 2009. New media vs. old media: A portrait of the Drudge Report 2002-2008. *First Monday 14*, 7.

[19] LEETARU, K. and ALTHAUS, S., 2009. Airbrushing History, American Style: The Mutability of Government Documents in the Digital Era. *D-Lib Magazine 15*, 1/2.

[20] MARKHAM, A.N., 2007. The Methods, Politics, and Ethics of Representations in Online Ethnography. In *Collecting and interpreting qualitative materials*, N. DENZIN Ed. SAGE, Thousand Oaks, CA, 247-284.

[21] MURPHY, J., HASHIM, N.H., and O'CONNOR, P., 2008. Take Me Back: Validating the Wayback Machine. *Journal of Computer Mediated Communication 13*, 60-75.

[22] PAYNE, N. and THELWALL, M., 2007. A longitudinal study of academic webs: Growth and stabilization. *Scientometrics 71*, 3, 523-539.

[23] ROBINS, G., PATTISON, P., KALISH, Y., and LUSHER, D., 2007. An introduction to exponential random graph (p*) models for social networks. *Social Networks 29*, 2, 19.

[24] RUMSEY, M., 2002. Runaway Train: Problems of Permanence, Accessibility, and Stability in the Use of Web Sources in Law Review Citations. *Law Library Journal 94*, 27-40.

[25] THELWALL, M. and VAUGHAN, L., 2004. A Fair History of the Web? Examining Country Balance in the Internet Archive. *Library & Information Science Research 26*, 2, 162-176.

[26] THELWALL, M. and VAUGHN, L., 2004. A fair history of the Web? Examining country balance in the Internet Archive. *Library and Information Science Research 26*, 2, 162-176.

[27] THELWALL, M. and WILKINSON, D., 2003. Three target document range metrics for university websites. *Journal of the American Society for Information Science and Technology 54*, 1, 29-38.

[28] VAUGHN, L. and THELWALL, M., 2003. Scholarly use of the Web: What are the key inducers of links to journal web sites? *Journal of the American Society for Information Science and Technology 54*, 1, 29-38.

[29] VERONIN, M.A., 2002. Where Are They Now? A Case Study of Health-related Web Site Attrition. *Journal of Medical Internet Research 4*, 2.

[30] WEBER, M.S., 2012. Newspapers and the Long-Term Implications of Hyperlinking. *Journal of Computer-Mediated Communication 17*, 2, 187-201.

[31] WEBER, M.S. and MONGE, P., 2011. The Flow of Digital News in a Network of Sources, Authorities, and Hubs. *Journal of Communication*.

[32] WEBER, M.S. and MONGE, P., 2014. Industries in turmoil: Driving transformation during periods of disruption. *Communication Research*, 1-30. DOI= http://dx.doi.org/10.1177/0093650213514601.