Web Archiving

Trevor Alvord
Brigham Young University, USA

INTRODUCTION

When the English philosopher Sir Francis Bacon first wrote *ipsa scientia potestas est* ("knowledge itself is power") in his book, *Meditationes Sacrae*, the most common way for knowledge to be disseminated was through the codex. By the 16th century, the printing press had transformed the codex and was in full use around the known modern world of the time. Through the printing press, knowledge and information became available on a mass scale to those who had not previously had access to it. Similar to Sir Bacon's era, an information revolution is taking place today through the Internet. Never before has human knowledge been so readily available. Yet, despite this great technological achievement, the Internet, like the codex, still suffers a potential type of rot. Information on the Internet is susceptible to loss, deterioration, or destruction. It is estimated that the Internet is decaying at a rate of 0.25% to 0.50% web pages per week (Fetterly et al., 2003). While a decay of less than 1% may not seem significant, the sheer size of the Internet—the website World Wide Web Size (2013) estimates that over 45 billion web pages had been indexed by Google during September 2013—translates into an estimated loss of over 112 million to 220 million web pages per week. Furthermore, the average life span of a web page is only about 100 days (Ashenfelder, 2011). This is a staggering amount of knowledge being lost on a daily basis. With so much knowledge rotting into the ether of cyberspace, a need for preserving the web is an acute task and parallels the same cultural need as preserving Bacon's *Meditationes Sacrae*. Nevertheless, recognizing this need is a rather simple process. Implementing a preservation strategy for web archiving with the integration of appropriate tools and standards can be much more daunting—but not impossible.

BACKGROUND

Jinfang Niu states (2012) that web archiving is "the process of gathering up data that has been recorded on the World Wide Web, storing it, ensuring the data is preserved in an archive, and making the collected data available for future research." (para. 1) Although Niu's excellent definition is expansive, current archival practice is more focused. Typically web archiving is done on the surface of the web, crawling and harvesting the HTML output that is seen by the user while avoiding the deeper, often larger files used to create that HTML output. These large sets of data or databases are typically being managed through data curation and not through web archiving.

Internet preservation began in 1996 with the formation of the Internet Archive, a non-profit organization founded by Brewster Kahle. In partnership with Alexa Internet, another Kahle-owned company that specialized in tracking web usage, crawling and capturing the World Wide Web began. In 1999, Amazon.com purchased Alexa Internet and Kahle started devoting more time to the Internet Archive. By 2001, Kahle launched the Wayback Machine (named for the WABAC Machine built by Mr. Peabody of the Rocky & Bullwinkle cartoon), which opened up to the general public the entire 10 billion URLs captured by the Internet Archive at the time (Green, 2002). The Internet Archive is no longer alone in preserving the web. Through its web crawling service Archive-it, the Internet Archive is adding around one billion pages per week and has archived over 360 billion URLs, all of which are available for use at waybackmachine.org (Rossi, 2013).

While the Internet Archive may have been the best known, it was not the only organization concerned about web archiving. The National Library of Australia also began an effort to capture and preserve websites under the .au domain, viewing websites as online publications and thus an extension of the national library's charge

DOI: 10.4018/978-1-4666-5888-2.ch757

to house a copy of Australia's published heritage. In order to accomplish this task, the Preserving and Accessing Networked Documentary Resources of Australia (PANDORA, http://pandora.nla.gov.au/) was launched in June 2001 (PANDORA: Australia's Web Archive, 2011). Similarly, the Library of Congress entered the fray with the launch of the web archiving project Mapping the Internet Electronic Resources Virtual Archive (MINERVA, http://loc.gov/minerva) in 2000. In 2004, the British Library started the UK Web Archive (http://www.webarchive.org.uk/ukwa). In 2005, the California Digital Library (CDL) received grant funding to create the Web-At-Risk Project. By 2006 the project had grown into the CDL Web Archiving Service (WAS). In the same year, Archive-It, a division of the Internet Archive, was created and began partnering with universities and organizations across the world for web archiving.

With a strong need for national libraries and government agencies to perform web archiving, the International Internet Preservation Consortium (IIPC) was founded in France in 2003. The consortium is membership driven, features five working groups, and is the leader in developing and promoting collaborative standards and tools for web archiving (International Internet Preservation Consortium, n.d.). Within the United States, there has been an absence of a national web archiving consortium until the Society of American Archivists (SAA) chartered their Web Archiving Roundtable in 2013.

ISSUES AND CHALLENGES

Archival Organization

A fundamental issue facing web archiving is administrative in nature; for instance, whether websites should be considered a published or unpublished work or whether web content represents a formal or informal record. As an organization approaches web archiving, these issues will fundamentally shape the direction taken. For instance, if an organization considers government officials' tweets to be public record, crawl cycles and retention schedules for preservation may be mandated at levels beyond the local organization's control. Additionally, particular tools or services for web archiving may not be able to match the specific retention cycle as mandated by law or by organizational policy. Providing access to harvested URLs will also potentially vary as websites and pages are identified as published or unpublished. For most library and archival organizations, published material is provided as a catalog record displayed through an Online Public Access Catalog (OPAC). Unpublished material tends to be grouped in collections and described in finding aids. Cataloging records verses finding aids may be more of an ideological debate; however, whichever record type is chosen will seriously impact resource allocations.

Appraisal

Archival appraisal is the process of determining if material has archival or historical value (Society of American Archivists, n.d.). Whether or not an organization is collecting internally for records management or externally for thematic material, archival appraisal of web content has to take place in real-time—before the information is lost. This becomes even more challenging for thematic web collections as the archivist or curator must determine on the fly, quickly and without being able carefully weigh it against established historiography, if a website or page holds enough value to justify archiving. In October 2013, the United States government shut down and, as a result, so did hundreds of government websites, each with its own unique shutdown statement. Three weeks later the shutdown ended and the sites came back up online. In this instance, only a three-week window of opportunity existed to appraise the value of creating such a collection. When the window passed, the likelihood of being able to recreate the content would be slight, if not impossible. Although this is an extreme example, it illustrates the challenges web archivists face as they begin to actively acquire web content.

Copyright and Ownership

One of the biggest challenges facing web archiving today is copyright and ownership. At present the prevailing interpretation of copyright for materials posted online is that every image, video, post, and comment is immediately copyrighted to the creator. This challenge is difficult enough for typical websites and blogs, but it becomes exacerbated in social media. There should be no doubt about the significance of social media.