Preservation of Web Materials

Julie Dietrich

INFO 560

Literature Review

7/20/13
Introduction

Websites are a communication and informational tool that can be shared and updated across the World Wide Web. Archiving web materials produces new and unique challenges for archivists, curators, and librarians. Websites are constantly changing and sometimes even disappearing at quick rates. Managing and capturing these materials into a collection is complicated. Collection development, methods, and collaboration are just some of the challenges discussed in the articles reviewed. Abbie Grotke’s “Web archiving at the Library of Congress” (2011), Rebecca Guenther and Leslie Myrick’s “Archiving web sites for preservation and access: MODS, METS, and MINERVA”, and Kathleen R. Murray and Inga K. Hsieh’s “Archiving web-published materials: A needs assessment of librarians, researchers, and content providers” all discuss the topic of archiving websites. While each article presented different issues, all authors agree that there is a definite need for preserving materials presented on websites for future generations.

Review of the Literature

Collection Development

Archiving web materials uses many of the same concepts as archiving traditional print documents. “Web collection development includes three major phases: selection, curation, and preservation” (Murray & Hsieh, 2008, p. 73). A theme, topic, or subject needs to be decided upon before capturing digital material. Types of websites might be another option in the selection process. Perhaps you only want government or educational websites in your collection. “Web archiving may not be covered by your organization’s existing policies” (Grotke, 2011, p. 16). Reevaluating one’s collection policy might be needed if web materials do not fall into the written purpose. Murray and Hsieh note concerns with collection development in their article
(2008). Librarians shared issues with selection and application of metadata through focus groups, interviews, and surveys. Reviewing materials requires more staff time and expertise during the collection development process which is another distress among professionals.

The scope of the project can be complicated. Archiving just a webpage would limit the research availability. Archiving a whole website seems to be the solution but that also involves looking at all internal and external links. The Library of Congress initiated a pilot project called the MINERVA (Mapping the Internet Electronic Resources Virtual Archive). They began by archiving websites related to the presidential election of 2000. These websites can change rapidly and capturing the historical evidence would have enduring value. The scope is another piece that should be decided during a collection development policy.

**Methods**

All articles state the same methods for capturing website materials: selective, harvesting, or deposit approaches. Web crawlers can assist in the process. The National Library of Sweden has chosen to use a harvesting approach by capturing the comprehensive domain of the country (Murray & Hsieh, 2008). Harvesting a whole country’s domain with limited staff resources and consideration of the scope of the project does not does work for most archivists. The Library of Congress (LC) has chosen to take a selective approach. “The LC uses the Heritrix web crawler—an open source crawler developed by the Internet Archive and national libraries and archives” (Grotke, 2011, p. 18). They have chosen to capture important historical events in there selective approach such as elections, 9/11, and Hurricane Katrina. These life-changing disasters that were shared over the World Wide Web are now archived for future generations to access.
Web crawlers do have their limits. Not all digital objects can be saved. “Heritrix is currently unable to archive streaming media, “deep web” or database content requiring user input” (Grotke, 2011, p. 18). Social media tools are another issue. Twitter and Facebook have been strongly used during elections and web crawlers have not anticipated these new tools. Archivists want to preserve the whole theme that has been presented on the World Wide Web. With technology constantly changing, there will always be new challenges present in capturing these websites.

Access

Saving these website materials for future use is just one piece of the process. Archivists must also make these available for patrons and researchers. The application of metadata can be challenging. In Murray and Hsieh’s article, “survey respondents identified cataloging as the top financial and technical challenge they anticipate in regard to building Web archives” (2008, p. 79). Limited staff resources and technical skills can be concerning in archival facilities and libraries. MARC records can limit viewing of the collection. “For the Election 2002 Web archive, LC collaborated with the SUNY Institute of Technology and WebArchivist.org for descriptive metadata for each Web site in the collection using the Metadata Object Description Schema (MODS)” (Guenther & Myrick, 2007, p. 148). By providing not only descriptive data but information such as ‘date captures’ and ‘access condition’, archivists are allowing users to see various captures of the selected website.

WebArchivist.org has developed a search system that has utilized MODS for users to be able to browse and search the database. The Library of Congress has chosen an open source software for access called Wayback Machine. Open source software can be an excellent way to save some money and test the waters of web archiving.
Collaboration

Collaboration and partnerships can be the key to managing the mass amount of data. “There is an opportunity at the state level for collaboration among state agencies, the state library, and university libraries” (Murray & Hsieh, 2008, p. 82). Partnerships could possibly help with cost and collaboration of expertise among library, archival, and information technology staff. The Library of Congress quickly tried to preserve the digital information surrounding the events of September 11th. “Partnering with Internet Archive and others enabled the collection of more than 30,000 websites during that time” (Grotke, 2011, p. 17). By creating partnerships, archivists can tackle the creation of a web archive.

Conclusion

“A Web site has to be one of the most complex and challenging of digital objects to capture, describe, manage, and preserve” (Guenther & Myrick, 2007, p. 143). Archiving materials from the World Wide Web presents many issues. Collection development policies, methods of capturing websites, and access to the information are some of the challenges that these articles have presented. Best practices have yet to be developed but are constantly changing with technology. Collaboration among facilities and staff can be a solution to the management of the overwhelming amount of materials. With so much communication and evidential information being shared on websites, preserving web materials for future users should be strongly considered by archivists, curators, and librarians.
References

