Against the Grain

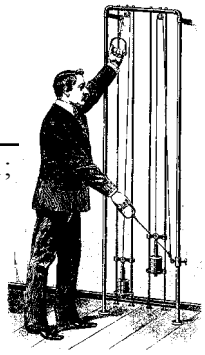
Volume 23 | Issue 1 Article 46

February 2011

Standards Column -- Preservation and the World Live Web

Todd Carpenter
NISO, tcarpenter@niso.org

Standards Column — Preservation and the World Live Web



by **Todd Carpenter** (Managing Director, NISO, One North Charles Street, Suite 1905, Baltimore, MD 21201; Phone: 301-654-2512; Fax: 410-685-5278) <tcarpenter@niso.org> *www.niso.org*

Our interactions online are increasingly based on customized profiles we set up on Websites, our past interactions on those sites, or our preferences, be they stated or computed. While we may visit sites like **Flickr**, **Yahoo**, **Facebook**, or our favorite online news site every day, we often don't realize that the page we see is very different from the Websites' homepages of a decade ago, or that they are very different from another user's experience. A decade ago, as the Web was just forming, it was certainly not a real-time experience, nor was it interactive. In 2003, **Allen Searls**, coined the term World Live Web (*http://blogs.law.harvard.edu/doc/2011/02/18/bring-on-the-live-web/*) as something that he envisioned the Web would become eventually — live and interactive in a way that it hadn't been before.

While we have not yet reached the state that the Web is live — most Websites still are primarily static and updated perhaps a few times per month — there are certainly aspects that are becoming more and more live. Anyone who followed the revolutions in Tunisia and Egypt through social media like **Twitter** and **Facebook** or through streaming media from news sites like **Al Jazeera** or **CNN**, there was an as-it-was-happening feel to the Web. They are examples of how, with a variety of syndication tools, easier to tweak and edit because of WYSIWYG editing tools, and trends toward more dynamically driven content, the Web is becoming increasingly interactive and real-time.

However there are downsides to this live interaction. Twenty-five years ago, television was a fairly live medium. VCRs could be used to record broadcast TV, but only if you knew in advance an event was taking place. And if you didn't happen to record the event, often the only recourse to get a copy would have been to contact the broadcaster and few if any of us did that. One of the problems of the live world is that it moves by pretty quickly and if we blink, we may miss it.

This was a topic of discussion during the LITA Top Tech Trends panel during the **ALA Midwinter** conference in San Diego in January. **Lorcan Dempsey**, Vice President and Chief Strategist at **OCLC**, made the point that we are all increasingly using a variety of social media, but there is little concern for or investment in the preservation of these repositories of content. While many of its users presume that **Flickr** will do a better job preserving our photos than they do, few question the underlying presumption that **Flickr** or any other Web 2.0 service will be available three, five, or ten years from now, certainly not that it will be available for my grandchildren — should I have any — to view my photos or other postings.

The permanence or preservation of live Web information is one of several critically important issues, especially for the library and scholarly communities. In a print world, there is usually a reference-able copy, even if in limited distribution and availability. In a digital world, the existence of the content continues only so long as the service provider doesn't replace the content with updates, deliberately delete content, fail to backup data, or even unplugs the service. There are a variety of ways that digital information can be changed, ranging from the innocuous (correcting errata) to the frustrating (lack of dates or version numbers that show changes have been made) to the malicious (attempting to clean up information about one's past behavior or limiting information about social protest movements). Content providers can easily change content in ways that impact the user experience through something as simple as changing the directory structure and the resulting URLs. Similarly, if each time we load a Web-based document, it is specific to our own profiles at the time and our experience is uniquely customized based on some aggregated or personal information, then there is truly nothing fixed, reference-able, or persistently linkable. Often even revisiting or reloading the same page could provide radically different information moments later. Some Websites even refresh automatically without any user interaction.

The scholarly journals community has made terrific strides with permanence of linking through the development and use of the Digital Object Identifier (DOI), (*http://www.doi.org*), which has seen tremendous success through applications such as OpenURL, DataCite, and others. Unfortunately, the problem becomes much more challenging outside the scholarly journals realm, where far fewer people think about this question or may not even consider it a problem. Some studies have pointed to a deterioration rate of URLs as being roughly 6-7% per year, which means that roughly half of the URLs noted in any given article will be non-functioning in about seven to eight years. Hopefully, with the completion of the standardization of the DOI System standard within ISO, the DOI will gain even broader adoption and acceptance as a solution to this problem. But that may require increased awareness of the problem outside of the academy or library community.

The **Library of Congress** has made some tentative steps in the direction of capturing this World Live Web through their acquisition of the entire dark-archive of **Twitter** last spring (*www.loc.gov/today/pr/2010/10-081.html*). While initially pilloried for their efforts, it was a prescient move on by **LC** to gather and archive this increasingly important public forum. Certainly not everything that is said on **Twitter** is worthy of preservation, however one doesn't know the gems until much later on in time and not archiving these items might be seen as a tremendous missed opportunity for the cost of several petabytes of storage space.

A second group that is engaging in finding solutions to this issue is a team led by **Herbert van de Sompel** from **Los Alamos National Laboratory** and **Michael Nelson** from **Old Dominion University**. They have been working on a project called Momento (*www.mementoweb.org/*). The prototype specification describes a system that allows users "to see a version of that resource as it existed at some date in the past, by entering that URL in your browser like you always do and by specifying the desired date in a browser plug-in." In December 2010, the **Institute for Conservation** and the **Digital Preservation Coalition** awarded Memento the **Digital Preservation Award 2010** (*www.dpconline.org/newsroom/latest-news/655-memento-project-wins-digital-preservation-award-2010*); it was also listed by the **Library of Congress** as one of the top ten technology achievements during 2010 (*www.digitalpreservation.gov/news/2010/20101229news_article_top-10stories.html*). The success of the project requires that the back-end content management system incorporate Memento into its service and maintain high-quality and consistent change log data. Wider adoption of Momento, especially by Live Web services, could prove an extremely useful tool in recreating the experiences of the past.

Also worth mentioning is the Atlas-like task underway by the Internet Archive of preserving as much of the open Web as possible. Unfortunately, the dynamically generated Live Web is less easily captured by the Archive's Wayback Machine (*web.archive.org*). Lesser-known work of the IA includes capturing and archiving of audio, concerts, and video. One of the biggest challenges the Internet Archive faces is how to deal with copyright. While preservation is an explicit exemption to copyright law in the United States, creating a closed archive for preservation purposes doesn't really serve a community where access is a crucial component of information curation.

There are undoubtedly benefits to having information as up-to-date as possible, and the instant communication opportunities and interactivity provided by a Live Web are tremendous. One need only look to the ever-changing situation in Middle East at the moment to understand the value. However, we need to adapt our existing structures and tools to manage this flexibility in a way that allows us to preserve the rapidly growing portion of our lives that we spend online. If we don't, the record of our digital-only content and interactions may fade as quickly as our memories of the live moment.