Brittan Dunham
Digital Preservation
Howard Besser
October 15, 2010

The Web-at-Risk: Preserving Our Nation's Digital Cultural Heritage

The ease of web publishing, access to information and collaborative grassroots efforts brought on by the Internet means there is more open discourse and information available to the public than ever before. Archivists understand that this wealth of information and points of view can be used to shape our national identity and compile a historical narrative, but face rapid technological advancements and a lack of resources and standards for preserving so much material. The California Digital Library understood that librarians needed a place to start, a framework within which they could begin to compile and preserve online information before the websites and their content were lost or no longer accessible.

The Web-at-Risk project began in 2005 when the National Digital Information Infrastructure and Preservation Program (NDIPP) awarded one of its first grants to the California Digital Library and its partners, New York University and the University of North Texas. The partners hoped to develop "tools to enable librarians and archivists to capture, curate, preserve, and provide access to web-based government and political information."[1] Additional support for the project has come from Stanford University, the San Diego Supercomputing Center, the Library of Congress and staff at several of the University of California library branches. This group of partners would be part of the experimental phase of the

[1] The Web-At-Risk: Preserving Our Nation's Digital Cultural Heritage

project, and would capture and curate a collection of websites for the initial launch of the service. If that proved successful, it would be made available to other libraries on a wider scale.

The majority of the work done for Web-at-Risk took place from 2005 to 2009 as the technical preservation infrastructure was developed. This infrastructure includes web-based applications that librarians and curators can use to create collections of online content and a digital repository where that content can be archived and accessed. As the partners planned how they would build the collection, they faced technical, procedural and cultural challenges. From a technical standpoint, they would have to develop an application that could compile large amounts of data quickly and handle the detailed user interface needed. The web-based application was eventually written using Java and Ruby on Rails, a web application framework. Procedurally they had to decide how to organize the steps to allow for the complicated procedure of creating a digital repository, while making the workflow as simple and user-friendly as possible. To help with this they developed a user manuals and detailed instructions so that libraries could use these tools or any similar web archiving infrastructure on their own in the future. There was also the cultural barrier of reaching out to traditionally document-centric libraries and developing a strategic plan for selecting websites that would be included in the collection. This phase of the project culminated with the Web Archiving Service (WAS) launch on July 15, 2009.[2]

[2] Digital Preservation News and Events

The WAS is a digital repository where the public can access the websites preserved by the partner libraries. The partners acted as curators, organizing captured sites by topic or thematic event, and harvesting specific sites for their relevancy and relation to the rest of the collection. The thematic focus was on preserving online information generated by state and local government organizations, citizen groups and political parties, though some web content from federal and international government organizations and non-profit organizations was also included. Partners were (and potential partners still are) encouraged to act as curators, building on and adding to these themes. The Web Archiving Service is now available to researchers and libraries beyond the University of California, but the service and repository, including the storage space and domain, remain hosted by California Digital Library.

In 2010, Web-at-Risk partnered with Ithaka Strategy and Research to gather information on the current state of web archiving and needs of librarians and archivists. They are looking at what has changed since the project began, how they can better meet the needs of libraries and archives, how they can reach out to a broader research community and what it will cost to sustain the application and the WAS. As it stands now, the public can access the WAS through the Californai Digital Library website at http://webarchives.cdlib.org/. The collections are listed there by theme or event and within each collection the archived sites are listed alphabetically by name, with links available for captures made on different dates. The site offers statistics about the archives and partners, information for researchers and the option for potential partners to begin using the service.

Sources

1. “The Web-At-Risk: Preserving Our Nation's Digital Cultural Heritage.” California Digital Library, 2010. http://www.cdlib.org/services/uc3/partners/webatrisk.html.

2. Murray, Katherine. Web-at-Risk Assessment Path Final Report. 2007.

3. Digital Preservation News and Events. “CDL Public Web Archive Service Collections Launched.” Library of Congress. July 15, 2009. http://digitalpreservation.gov/news/2009/20090715_article_was.html.

4. Digital Preservation Partner Tools and Services: Web at Risk. Library of Congress. http://digitalpreservation.gov/partners/resources/tools/index.html#w.

5. Web Archiving Service. http://webarchives.cdlib.org/.