

Digital Archiving: Making it Happen

Defining a Web Archiving Strategy

Kristine Hanna

Internet Archive
kristine@archive.org

Kristine Hanna, Director of Web Archiving Services at Internet Archive, will discuss the significance of web archiving, the challenges libraries, archives and memory institutions face in the digital age, as well as some of tools and best practices currently in use to create a successful web archiving strategy.

The Internet Archive, located at <http://www.archive.org/index.php> on the web, has been involved in web archiving since 1996 when the organization was founded as an internet library to provide permanent access for researchers, historians and the general public to the world's cultural artifacts. Additionally the web group at the Archive <http://wa.archive.org/> works with institutions to create focused collections through crawling services and Archive-It, a web based application.

The Internet Archive is a founding member of the IIPC (International Internet Preservation Consortium <http://netpreserve.org/about/index.php>), and we work closely with national libraries and archives from around the globe to develop open source tools and document best practices for web archiving.

Libraries and archives have long collected information that serves scholars, historians and the general public in understanding history, culture, and society. So much of today's information is easily found on the world wide web – web pages have replaced hard copy newsletters, blogs are today's diaries, many government forms and documents are more readily accessible on the web than they are in paper form. And there is one sure thing about the World Wide Web: like the weather, it will change. An estimated 44 percent of Web sites that existed in 1998 vanished without a trace within just one year. The average life span of a Web site is only 44 to 75 days.

With the rapid growth of the Internet and the Web, hundreds of millions of people around the world have grown accustomed to using the web as their primary resource to acquire information. The availability of this electronic information is taken for granted. It is a fallacy that if something is on the web it will be there forever. There's an urgent need for people to understand that the web is who we are. It's our culture and our social fabric, and we don't want to lose any of it. What is here today might be gone tomorrow.

As part of an effort to appropriately document and capture today's information for tomorrow's use, institutions must adopt a web archiving strategy. However, for many institutions, the prospect of capturing and storing web sites or entire web domains is a daunting prospect. This presentation will include an overview of web archiving at Internet Archive with our 60+ partners in national, state, university and public libraries. We'll take a look at why our partners are invested in web archiving, how they are applying web archiving in their particular institutions, some of the challenges

they face, what benefits they have seen, the tools they use, their lessons learned, as well as what they are looking forward to in the future.

We'll also take a look at "Best Practices" around web archiving, what our partners have uncovered, what they are recommending, what we have learned. We'll review some of the questions partners ask when planning a digital archiving strategy. What are the biggest challenges your institution faces in archiving and preserving digital materials? What is prompting your institution to look for a web archiving service? Is there a specific reason or event? How is your institution using the collections you have gathered?

Lastly, we'll provide an overview of the digital stewardship role that the Internet Archive is playing re: archiving and preserving data from around the world, with our hopes and objectives for the future, including the roles that other international institutions could play in furthering the mission.

Internet Archive is a member of the American Library Association, Society of American Archivists and is officially recognized by the State of California as a library. All collections are publicly accessible with free access at www.archive.org.

The Internet Archive is a non-profit organization founded in 1996 as an internet library to provide permanent access for researchers, historians and the general public to the world's cultural artifacts. The Archive collaborates with federal and international institutions including the U.S. Library of Congress, National Library of Australia, the Bibliothèque Nationale de France and the Smithsonian, plus over 45 national, state, university and public libraries and archives. The Internet Archive has the largest public web archive in existence with over 2 petabytes of data containing 85 billion pages from over 65 million websites in 37 languages. The Internet Archive recently shipped 1.5 petabytes of data to our partner, Bibliotheca Alexandrina, in Egypt; a leading institution of the digital age and a center for learning, tolerance, dialogue and understanding.