Date : 26/05/2008

Preserving Access to Government Websites: Development and Practice in the CyberCemetery

Starr Hoffman
University of North Texas Libraries,
Denton, Texas
United States of America

Meeting: **130. Government Information and Official Publications**
Simultaneous Interpretation: English, Arabic, Chinese, French, German, Russian and Spanish

WORLD LIBRARY AND INFORMATION CONGRESS: 74TH IFLA GENERAL CONFERENCE AND COUNCIL
10-14 August 2008, Québec, Canada
http://www.ifla.org/iv/ifla74/index.htm

ABSTRACT:

In the late 1990's, online U.S. government information was appearing and disappearing at a rapid pace. In 1999, the University of North Texas Libraries (UNT) formed a partnership with the U.S. Government Printing Office (GPO) to address this issue by archiving electronic government websites. This archive, known as the CyberCemetery, provides permanent public access to the websites and publications of defunct U.S. government agencies and commissions. This partnership between UNT and GPO has expanded to include the National Archives and Records Administration (NARA). This paper covers the CyberCemetery's development and the process of identifying, capturing, and publishing content in the archive.

FULL TEXT:

Introduction

In the late 1990's, online U.S. government information was appearing and disappearing at a rapid pace. This information was often born-digital with no hard-copy backups, meaning that if it was edited or removed, the originals were lost. In 1999, the University of North Texas (UNT) Libraries formed a partnership with the U.S. Government Printing Office (GPO) to address this issue by archiving electronic government websites.

This archive, known as the CyberCemetery, provides permanent public access to the websites and publications of defunct U.S. government agencies and commissions. The collection is housed on servers at the University of North Texas (UNT) Libraries and is available online at http://govinfo.library.unt.edu. The archive features information on a variety of topics and such important records as the website of the National Commission on Terrorist Attacks Upon the United States (commonly referred to as the 9/11 Commission). The CyberCemetery

has grown from a single website in 1997 to forty-five in 2008, with nine more in the process of being archived. Each month, it receives an average of 220,000 visits and one million hits.

Users can access the CyberCemetery online from any location, free of charge. This satisfies the Federal Depository Library Program (FDLP)'s mission to provide permanent public access within FDLP libraries, and reaches beyond that mission to provide global access to this valuable material. Although 79% of users access this archive within the United States, there is a significant minority of access from regions such as Europe (5.67%) and Asia (3.77%). The largest groups of non-U.S. users by country are China (1.6%), Canada (1.5%), and the United Kingdom (1.2%).

It is important that this content remains accessible because of its immense value. The CyberCemetery features a variety of topics indicative of the broad nature of government information. The archived websites contain statistics, commission reports, policy recommendations, photographs, videos and transcripts from hearings, and other documents. This information is valuable for government employees, for the American public, and for the global community. In addition, it is important to preserve these documents for the future as part of the American historical record.

Development of the Partnership

In 1995, the U.S. Government Printing Office (GPO) published a draft report of their strategic plan, which emphasized the need to preserve electronic agency publications. The GPO and the depository library community were concerned that government agencies were removing information from their websites without regard to permanent preservation. In GPO's 1996 strategic plan, *Study to Identify Measures Necessary for a Successful Transition to a More Electronic Federal Depository Library Program*, it called for the formulation of partnerships with depository libraries to identify and archive this material[i].

Cathy Hartman, a government documents librarian at UNT, initiated discussion with GPO about forming such a partnership. The initial collection selected for preservation was the website for the Advisory Commission on Intergovernmental Relations (ACIR), a defunct government agency. Because the agency was defunct, no more content would be added, and the website was to be taken offline. This website was selected for preservation both because of its at-risk status and its static content, which would keep preservation and storage costs low. The original Memorandum of Understanding (MOU) developed between GPO and UNT specified that the primary purpose of this project was to provide free, unrestricted, and permanent access to the agency's website. Many of the stipulations of the MOU echoed the traditional guidelines of the Federal Depository Library Program (FDLP), which provides no-fee, permanent access to print resources.

This agreement was expanded in 1999 to include additional websites, all from government agencies or commissions that have ceased operation. This revision came about because of the urgent need for preservation of this material. Many of the documents on these websites is delivered only electronically; therefore when the agency or commission hosting those documents closes and the website is removed, there is no longer any public access to those documents. After the archive was expanded it became known as the CyberCemetery, indicating its role as an online archive for defunct government agencies and commissions.

This partnership between GPO and UNT was modified again in 2006 to include the National Archives and Records Administration (NARA). UNT is now an Affiliated Archive of the National Archives, making it one of only three higher education institutions to receive that distinction. Once websites are archived in the CyberCemetery, NARA accessions the records as part of the Archives of the United States.

Archive Scope

As stated previously, the CyberCemetery's scope is any website of a now-defunct government agency or commission. Once we indentify an agency or commission whose termination is imminent, we watch the website closely. We capture it only once, archiving its most complete and final version. The capture is initiated when the agency becomes inactive, when an expiration date has passed, or after a final report has been published. The complete website is archived, including files of all types.

Archiving Process

The process of archiving materials in the CyberCemetery has evolved over the past decade. In the current process, UNT first identifies at-risk government agencies or commissions. I usually perform this as part of my duties as the Librarian for Digital Collections in the Government Documents Department. However, I also receive website referrals from other librarians in the FDLP community. As the CyberCemetery has become well-known, some government agencies and commissions directly contact GPO or UNT to inquire about the possibility of archiving their website.

Next, the website is evaluated. It must meet the following criteria:

1) It must be the official website for a federal government agency or commission.
2) The government agency or commission must be closing, have issued a final report, or otherwise indicate that the website is at-risk.

If the agency or commission has directly communicated a desire to have the website archived, we also send the following questions to their contact or web administrator:

- What operating system was used to host this website?
- What webserver software was used for hosting of this website?
- Are server side includes (ssi) used in this website?
- Was this website static html or a dynamic site?
    - If so, what scripting languages were used for this website (php, perl, python)?
    - Was a database used for this website?
        - If so, what database was used for this website?
        - If so, what methods were used to connect to the database?
- Is there streaming media associated with this website?
- Are there proprietary content types used in this website?
- Are there any comments you would like to add?

These questions help us determine if the website can be easily and completely captured by web harvesting. Our past practice was to harvest websites using HTTrack, a free web harvesting application (http://www.httrack.com/page/1/en/index.html). This software allows you to download an entire website including HTML, images, and other file types.

We are currently transitioning from HTTrack to use Heritrix as our harvesting software (http://crawler.archive.org/). Heritrix is an open-source web crawling application that was developed for the Internet Archive. Heritrix archives websites into ARC files, which store multiple archived resources in a single file that ranges from 100 to 600MB in size. Once we have developed an interface for these files, we will be able to display and indicate the websites' archival status without altering the original code. This will be an improvement over the previous method, which required minor changes to disable contact pages and to indicate the archival status.

On occasion, websites have been obtained when the agency or commission has donated the content. This entails copying the files to portable media and mailing that media to UNT. This was done in the instance of the Office of Technology Assessment (OTA), the 9-11 Commission, and others. In the future, this will be done only in cases when the website contains content that is not obtainable by harvesting. Harvesting provides greater assurance that files have not been altered from their original state as published online at the time of the agency or commission's termination.

At this time, we are not constricted by the size of the website to be archived. We currently have 13GB of server space available for new content. When more is needed, the Government Documents and Information Technology Services (ITS) departments will discuss the need and possible solutions. We do not anticipate problems for future expansion, as storage space is relatively inexpensive at this point.

When the entire site has been harvested or otherwise received, it is checked for errors and completeness. This is both a manual and automated process. The archived site is manually navigated and compared with the original. We also run link-checking software called Xenu Link Checker to produce a report of any broken links in both the archive and the original site (http://home.snafu.de/tilman/xenulink.html). We then compare the results of both reports to identify any files that have been overlooked and need to be captured before the archived website goes live. Any overlooked files are harvested from the original site, if possible. Other options include capturing the file from the Internet Archive, or contacting the agency or commission through GPO.

The Memorandum of Understanding (MOU) between UNT and GPO allows slight alteration to an archived website's code in order to indicate that it is no longer live. Standard practice has been to mark each page of an archived website for this purpose. To meet the requirements of changing the original code as little as possible, the agreement permits adding the text "Archive" in 8 point, Times New Roman font at the top of each page. To meet the CyberCemetery's goals of preservation both in terms of accessibility and authority, our new process of harvesting websites using Heritrix will enable us to indicate the websites' archival nature without altering the code. In addition, it will save us from the lengthy process of altering every page of each archived website.

Our original practice was to disable contact links, such as email address links, as these are no longer current or relevant once an agency or commission has closed. However, when

Heritrix is regularly used to harvest these websites, no changes will be made to the code. The CyberCemetery homepage will indicate the archival nature of the websites and the inoperability of any contact information or links.

Next, the website is loaded onto UNT's web server. We also add a link on the CyberCemetery "Browse" pages, which are arranged by title, by branch of government, and by date of expiration. If GPO or the agency or commission initiated contact about the harvest, we notify them at this time that the site is live.

When a government commission has self-selected for inclusion in the CyberCemetery, they sometimes wish to keep their original domain name active for a period of time after the archived website goes live. In this case, we provide the commission (or GPO, in the case that GPO is the commission's acting web administrator) with the archived website's Internet Protocol (IP) address. It is the government commission or GPO's responsibility to extend their domain name subscription and to create the redirect action from that page to the CyberCemetery's archived version. This helps the public find the website during the transition period, during which the old Uniform Resource Locator (URL) will appear near the top of search engine results for that group. After the archived version has been up for some time, it tends to appear near the top of most search engine results, and the redirect is no longer necessary. Thus far, our experience has been with groups extending their domain name subscriptions for one to two years for this purpose.

After the archived site is live, we incorporate it into the UNT Libraries' Digital Library System (DLS) (http://digital.library.unt.edu). This is the online system that incorporates all of UNT's digitized collections, as well as the born-digital government documents and websites that we capture. Each archived site on the CyberCemetery is recorded at the site-level in the DLS to provide additional access to this content. The websites are subject-classified using the Legislative Indexing Vocabulary (LIV), which was created by the Congressional Research Service (CRS) to describe legislative and public policy literature. These records are useful discovery tools because they help identify relevant CyberCemetery sources to UNT patrons who may be familiar with UNT's general digital collections, but unaware of the CyberCemetery by name. The DLS enables patrons to perform searches in a single interface that includes websites, images, documents, and multimedia across a variety of subjects and collections. Finally, these DLS records enable us to share metadata about the CyberCemetery with other institutions through the Open Archives Initiative (OAI) and federated searching, increasing the public's access to and awareness of this content.

To add a record to the DLS, I first create a thumbnail image from the website's homepage. I do this by viewing the website in my browser and pressing the "print screen" key. I then open an image editing application such as Adobe Photoshop and paste the image into a new file. I crop the image and resize it, generally to 118 x 90 pixels at about 100 dpi (dots per inch). The resulting thumbnail image is usually less than 5k (kilobytes).

I then create a metadata record at the site-level; the metadata used for this collection is based on the UNT Libraries metadata scheme (http://www.library.unt.edu/digitalprojects/metadata). One of the fields is a description of the content of the website. This is a three to four-sentence paragraph that describes the website and includes keywords that increase the relevancy of the DLS's search results. This record is linked to the full archived website on the CyberCemetery server. When the record is complete,

I notify our Digital Projects Unit that the metadata record and icon are ready to be uploaded to the DLS.

Equipment, Environment, & Backup

To ensure permanent access to these websites, the CyberCemetery servers are housed in a controlled environment in the library basement. The server room is kept at 38 degrees Fahrenheit (about 3 degrees Celsius) and 50% humidity, on average. The CyberCemetery is housed in a four node fail-over clustered configuration using a SAN volume for storage. The CyberCemetery is active on only one server at a given time, while the other three servers act as backups in case of hardware or software failure or maintenance. Currently the CyberCemetery encompasses 27.2GB of content on a 40GB volume server. Filesystem storage allocation is discussed between our Government Documents and Information Technology Services (ITS) departments. On weekends, full backups are made to magnetic tape, which is then shipped to off-site storage at a company called Iron Mountain (http://www.ironmountain.com).

Data Migration

Some file types on the CyberCemetery are unique and proprietary. They may become difficult to use or read newer software versions, or the current software may become completely obsolete. These file types do not always have adequate documentation on how to access the content contained within the file. Because the primary purpose of the CyberCemetery is to preserve public access and usability, as files become unusable they will be migrated to a usable format.

To this end, we will perform a site inventory of file types currently on the CyberCemetery. We will then identify which of these file formats are at-risk and investigate and document a file migration plan for those formats. This plan will consider the important issues of both functional access and preservation of the original look and feel of the content. The formation of this plan will include the involvement of the Superintendent of Documents staff, as specified in the original Memorandum of Understanding (MOU) between GPO and UNT. The at-risk files will be migrated to new formats as becomes necessary to preserve access. The original files will be retained on the CyberCemetery server to permanently preserve them. This is acceptable as per the MOU in order to retain accessibility to this content. All changes of this type will be documented at the website level.

Conclusion

The CyberCemetery was formed to provide permanent public access to government information that is disappearing from the internet. It has transformed from a single website to forty-five websites maintained by a partnership between the U.S. Government Printing Office, the National Archives and Records Administration, and the University of North Texas Libraries. As technology continues to evolve, we will seek further methods of preserving this information and maintaining its usability. As a member of the Federal Depository Library Program, UNT is committed to permanent public access to this varied and valuable information.

[i] Hartman, C. N. (2000). Storage of electronic files of federal agencies that have ceased operation: A partnership for permanent access. *Government Information Quarterly*. *17*, 299-307.