3-2012

Functionalities of Web Archives

Authors: Jinfang Niu

The functionalities that are important to the users of web archives range from basic searching and browsing to advanced personalized and customized services, data mining, and website reconstruction. The author examined ten of the most established English language web archives to determine which functionalities each of the archives supported, and how they compared. A functionality checklist was designed, based on use cases created by the International Internet Preservation Consortium (IIPC), and the findings of two related user studies. The functionality review was conducted, along with a comprehensive literature review of web archiving methods, in preparation for the development of a web archiving course for Library and Information School students. This paper describes the functionalities used in the checklist, the extent to which those functionalities are implemented by the various archives, and discusses the author's findings.

Scholar Commons Citation

Niu, Jinfang, "Functionalities of Web Archives" (2012). *School of Information Faculty Publications*. 309.
http://scholarcommons.usf.edu/si_facpub/309

D-Lib Magazine

March/April 2012
Volume 18, Number 3/4

Functionalities of Web Archives

Jinfang Niu
University of South Florida
jinfang@usf.edu

doi:10.1045/march2012-niu2

Abstract

The functionalities that are important to the users of web archives range from basic searching and browsing to advanced personalized and customized services, data mining, and website reconstruction. The author examined ten of the most established English language web archives to determine which functionalities each of the archives supported, and how they compared. A functionality checklist was designed, based on use cases created by the International Internet Preservation Consortium (IIPC), and the findings of two related user studies. The functionality review was conducted, along with a comprehensive literature review of web archiving methods, in preparation for the development of a web archiving course for Library and Information School students. This paper describes the functionalities used in the checklist, the extent to which those functionalities are implemented by the various archives, and discusses the author's findings.

Keywords: web archive, functionality, usability, evaluation, overview

Introduction

In order to preserve human cultural heritage and to keep continuity of access to content that has already disappeared from the web, or soon will disappear, many libraries and archives have started to archive web resources. Much has been reported about how libraries and archives select, acquire, store, organize and describe web resources for preservation. As libraries and archives become more settled with these collection-building practices, they need to pay more attention to the users, usability, and functionality of web archives, since the ultimate purpose of archiving web resources is to use them.

There have been a limited number of studies on the uses and functionalities of web archives. The International Internet Preservation Consortium (IIPC) Access Group has defined a variety of use cases that web archives are expected to support (IIPC Access Working Group, 2006). Each of these use cases requires a number of functionalities. Many of these functionalities, such as "search by URL", are basic to search and retrieval in web archives. A number of functionalities, such as "data mining," require more advanced means of information discovery and analysis. The IIPC use cases can be used as best practice guidelines for the functionalities of web archives.

After the publication of the IIPC use cases, two user studies identified user expected or preferred functionalities. Ras & Bussel (2007) studied user needs and identified potential types of users of web archives of the National Library of the Netherlands (KB). Based on their user study, Ras & Bussel identified user-preferred search interface features of web archives, which include both usability and functionality features.

The most recent user study on web archives was conducted by Costa & Silva (2010) on the Portuguese Web Archive (PWA). They studied the needs and behaviors of people searching the PWA. They found that searching for a known page or site was more frequent than collecting information about a subject written about in the past, which was more frequent than performing web-based transactions. Another finding is that users prefer the oldest documents over the newest. This study also identified user preferred or expected functionalities. For example, although users prefer full-text search over URL searches, URL queries are common and should be supported. Specialized search engines for images, videos and old news, and the ability to demonstrate the evolution of a page or site, are also expected.

To expand upon the findings of a comprehensive literature review of web archiving methods, the author conducted an overview of the functionalities supported by current web archives, by creating a checklist of functionalities based on the IIPC use cases and the two user studies mentioned above. This checklist, found at Appendix 1, was used as a benchmark to evaluate the functionalities of ten member web archives of IIPC. Although the sample size is small due to language barriers, and the access restrictions of some web archives, the findings did reveal the current functionalities provided by some of the most established web archives in the English speaking world (or for which there is an English language interface). The findings also highlight the gap between the current functionalities, and the expected functionalities identified by the IIPC use cases and recent user studies. The checklist and current findings may help to inform the functionality design of future web archives and assist future archivists. The checklist can also be used as an evaluation or self-assessment tool for existing web archives.

Methodology

Web archives evaluated

This project evaluated the functionalities of ten member archives of IIPC. IIPC has a web archive registry, netpreserve.org (International Internet Preservation Consortium), that includes 24 web archives. The archive of one of the 24, Bibliotheca Alexandrina, is a mirror site of the Internet Archive (IA), and provides identical functionalities to the IA Wayback Machine. Since IA is also on the registry, Bibliotheca Alexandrina's web archive was excluded from the study. Eight of the archives are either dark archives or allow only onsite access, making them inaccessible.

Fourteen web archives are publicly accessible online; however, only nine of them, listed below, have an English interface:

- IA Wayback Machine (started in 1996)
- Preserving and Accessing Networked Documentary Resources of Australia (PANDORA) Web Archive created by the National Library of Australia (started in 1996)
- UK Government Web Archive created by the UK National Archives (started in 1997)
- New Zealand Web Archive of the National Library of New Zealand (started in 1999)
- Library of Congress Web Archives (started in 2000)
- Web Archiving Service (WAS) of the California Digital Library(started in 2003)
- Government of Canada Web Archive created by Library and Archives Canada (started in 2005)
- Web Archive Collection Service (WAX) created by Harvard University Library (launched in 2009, piloted in 2006)
- UK Web Archive provided by the British Library in partnership with the National Library of Wales, Joint Information Systems Committee and the Wellcome Library (started in 2005)

The Archive-It service (started in 2005) is provided by IA and is not listed as a separate member on the IIPC registry. However, Archive-It and the IA Wayback Machine provide very different functionalities. They were included as two separate web archives in this study, bringing the total to ten.

These ten web archives vary in their collection scope, size and age. At the time of the study, IA Wayback Machine was 15 years old, but WAX was only two years old. The collection scope of IA Wayback Machine is the most comprehensive, followed by the web archives of several national libraries. The UK Government Web Archive and Government of Canada Web Archive preserve only government websites. WAX, due to its very young age and the fact that it is provided by a university library, only had four collections at the time of this study.

Unlike the other eight web archives, WAS and Archive-It are service providers that provide technical infrastructure, data storage and training for other organizations. Subscribers of WAS and Archive-It can focus on resource management issues such as organization, description and access control, and do not need to worry about the technical concerns of web archiving. These two service providers also represent the web archiving practices of a large number of organizations in the United States. As of April 2011, WAS had 16 subscribers and Archive-It had over 160 subscribers, including national libraries, state libraries and archives, academic libraries and government agencies. Many subscribers provide a link on their websites to Archive-It and do not have their own web archive interface, such as the Arizona State Library, Archives and Public Records and the Michigan Government Web Collection.

Although ten web archives sounds like a small sample size, their status as member archives of IIPC shows that they are among the most established web archives in the world. In addition, since many subscribers rely on WAS and Archive-It to provide an access interface, the functionalities of these two web archives also represent the functionalities of the web archive collections created by many other organizations.

Creation of the functionality checklist

The IIPC use cases were divided into five categories. The "Use Cases for Archive Internal Use" in the last category were excluded from this study because they were created for people who build the web archives. Content analysis was conducted on the first four categories of the use cases. Functionalities required to support each use case were identified. For example, one use case states:

> *"Jane Jones learns that a competitor of her company copied the appearance, including trademarks, of her business website, and engaged in a mass mailing to customers directing them to the spoofed website, where some were tricked into supplying proprietary information. By the time legal action is contemplated, the competitor has removed the offending material from the website, but copies exist in the ArcSys. Jane Jones keys in the URL and relevant dates of the offending website, and wishes to receive evidentiary-quality printouts of the relevant pages, which may require signed declarations of ArcSys personnel."* (IIPC Access Working Group, 2006, p. 8)

This use case requires the following functionalities: search by URL, record the dates on which the URL was harvested, allow people to print out authentic copies and certify the authenticity of a printed copy.

Some functionalities, such as search by URL, were identified from more than one use case. The duplicates were removed and all the remaining functionalities identified from the use cases were gathered together, and combined with the user-preferred or expected functionalities extracted from the two users studies (Ras & Bussel, 2007; Costa & Silva, 2010). There was also overlap in the functionalities identified from the two user studies and those from the use cases. Again, duplicates were discarded. In the end, all the remaining functionalities were re-organized into groups: search parameters, search results, browsing, policy-related functionalities, personalized services, data mining, and reconstruction of lost websites. The final functionality checklist found at Appendix 1 was thus created.

Method of investigation

The ten archives were evaluated using the checklist in April 2011. The author mined the interfaces of targeted web archives thoroughly to find out whether each functionality was supported by each web archive. This included reading information about the archives (Frequently Asked Questions, technical information, statistics reports, etc.), conducting searches, observing the ranking and presentation of search results and the way un-archived web pages were dealt with, trying to print the web pages, and testing whether the forms and search boxes work in some web archives. Based on this exploration, the author judged whether each functionality was supported, and then recorded the findings in a table. The results were either "Yes," "No," or "No information found", occasionally supplemented by notes about how the functionality was supported. The author also recorded some functionalities provided by the web archives but not covered by the checklist, and some evident usability issues discovered along the way. One item on the checklist, "Archived websites are the same as the original version of this site" was not used in the evaluation because the original versions of many archived websites are no longer available on the Internet.

Findings and Discussion

Search parameters

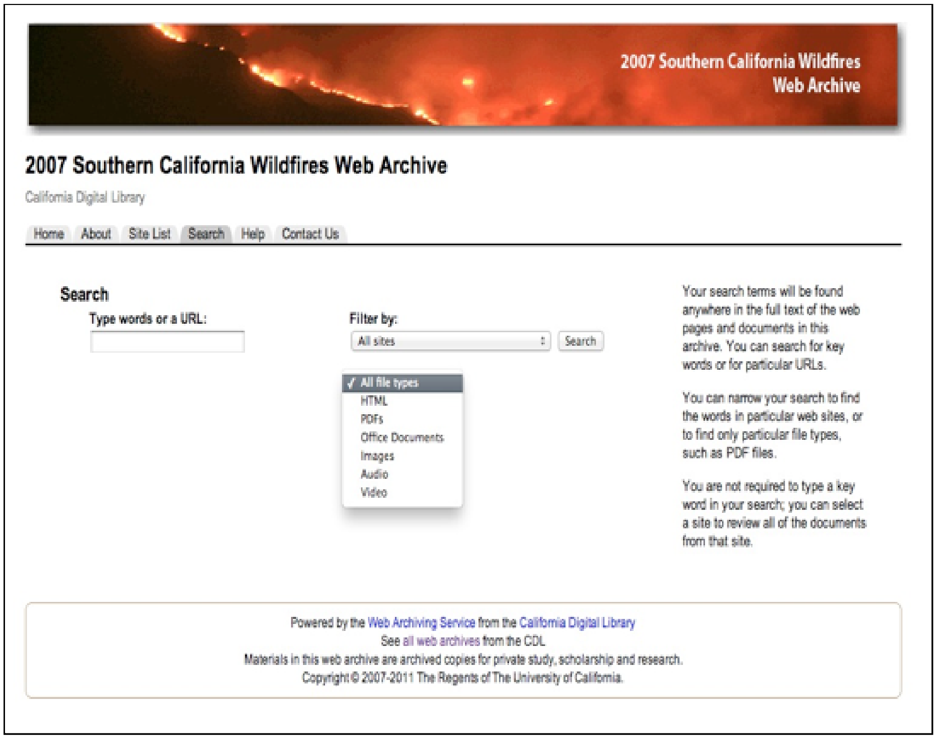
> ***Search by URL and Keyword:*** URL is the mostly widely supported search parameter followed by keyword. Six web archives support search by exact URL. The other four support search by URL as a keyword. Searching by exact URL and searching using a URL as a keyword sometimes return different search results. Searching the URL "www.umich.edu" at IA Wayback Machine will only return web pages at that URL, whereas searching using that

URL as a keyword at Archive-It will return results that contain the URL anywhere in the archived web content, for example, "http://www.ischool.utoronto.ca/resources/inforum/blog/2005/04/www.umich.edu" (contains www.umich.edu in the URL) or "http://www.michigan.gov/dmva" (contains www.umich.edu on the web page). The Library of Congress Web Archives and the New Zealand Web Archive support keyword search on bibliographic records. The other web archives support keyword search on the full text of web pages.

Domain-based search: Five web archives support restricting a search by domain, offering users the option of searching content within a single website. PANDORA allows a search to be limited to a given Top Level Domain, e.g., .gov or .edu.

Narrow the search by date: Six web archives support narrowing searches by date, and the granularity of the dates varies. The Government of Canada Web Archive, IA Wayback Machine and UK Web Archive allow the user to limit the search to a time period defined by days (YY/MM/DD - YY/MM/DD). Archive-It can limit the search to a time period defined by months (YY/MM - YY/MM). PANDORA allows its users to limit the search to a time period defined by years (YY-YY). The Library of Congress allows the user to restrict the search to a particular year. Theoretically, larger web archives need more detailed granularity to avoid very long lists of search results. It is not known how these web archives decide the granularity nor whether the size of the web archives plays a role in their decision.

Narrow search by media type: Seven web archives support searching by media type. The media type list provided by the web archives varies. The UK Government Web Archive lists only two types: HTML and PDF, whereas the UK Web Archive lists eight media types: HTML, PDF, MS Word, MS Excel, PPT, images, video and audio. This difference might be caused by the differences in the web archives, as the UK Web Archive is broader and may contain more media types. The web archives also categorized media types differently. For example, IA Wayback Machine uses "text" as the media type, while Government of Canada lists specific types of text: 'DOC' and 'HTML'. The UK Web Archive lists 'MS Word', 'MS Excel' and 'PPT', whereas WAS uses 'Office Documents', which covers all three types in the UK Web Archive. In most web archives, the search by media type feature is provided on the search interface as a drop-down box and serves as a pre-search filter, as shown in Figure 1 below.


2007 Southern California Wildfires Web Archive

2007 Southern California Wildfires Web Archive

California Digital Library

Your search terms will be found anywhere in the full text of the web pages and documents in this archive. You can search for key words or for particular URLs.

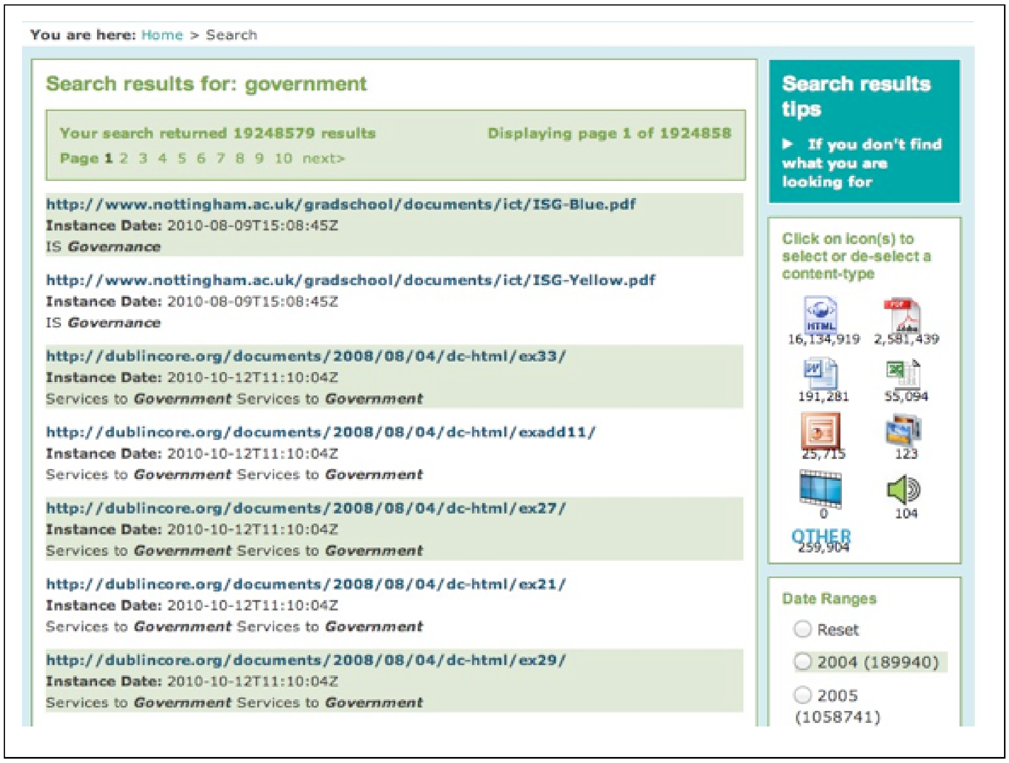
You can narrow your search to find the words in particular web sites, or to find only particular file types, such as PDF files.

You are not required to type a key word in your search; you can select a site to review all of the documents from that site.

Powered by the Web Archiving Service from the California Digital Library
See all web archives from the CDL
Materials in this web archive are archived copies for private study, scholarship and research.
Copyright © 2007-2011 The Regents of The University of California.

Figure 1: Media type as pre-search filter.

The UK Web Archive presents this feature as icons on the search results page and uses them as a post-search filter. When a user clicks on an icon, the search results are filtered by the media type represented by that icon, as shown in Figure 2.



You are here: Home > Search

Search results for: government

Your search returned 19248579 results — Displaying page 1 of 1924858

Page 1 2 3 4 5 6 7 8 9 10 next>

http://www.nottingham.ac.uk/gradschool/documents/ict/ISG-Blue.pdf
Instance Date: 2010-08-09T15:08:45Z
IS *Governance*

http://www.nottingham.ac.uk/gradschool/documents/ict/ISG-Yellow.pdf
Instance Date: 2010-08-09T15:08:45Z
IS *Governance*

http://dublincore.org/documents/2008/08/04/dc-html/ex33/
Instance Date: 2010-10-12T11:10:04Z
Services to *Government* Services to *Government*

http://dublincore.org/documents/2008/08/04/dc-html/exadd11/
Instance Date: 2010-10-12T11:10:04Z
Services to *Government* Services to *Government*

http://dublincore.org/documents/2008/08/04/dc-html/ex27/
Instance Date: 2010-10-12T11:10:04Z
Services to *Government* Services to *Government*

http://dublincore.org/documents/2008/08/04/dc-html/ex21/
Instance Date: 2010-10-12T11:10:04Z
Services to *Government* Services to *Government*

http://dublincore.org/documents/2008/08/04/dc-html/ex29/
Instance Date: 2010-10-12T11:10:04Z
Services to *Government* Services to *Government*

Search results tips

▶ If you don't find what you are looking for

Click on icon(s) to select or de-select a content-type

Date Ranges

○ Reset

○ 2004 (189940)

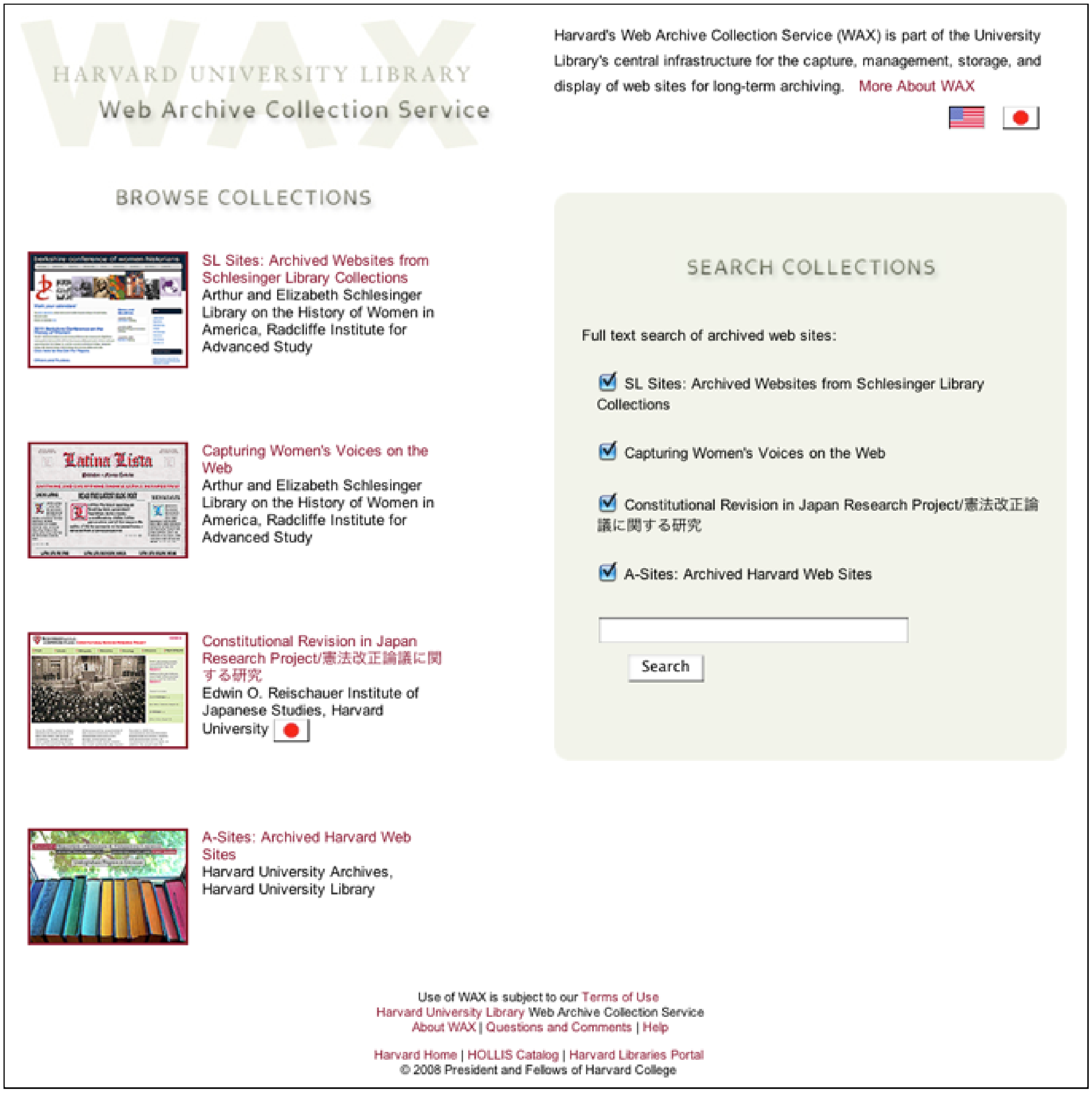
○ 2005 (1058741)

Figure 2: Media type as post-search filter.

Integrated search: All web archives that are affiliated with a traditional library are accessible through the library catalog (the New Zealand Web Archive, PANDORA, WAX and the Library of Congress Web Archives). Archive-It provides an integrated search interface for the web archive collections of all its subscribers. WAS provides separate search interfaces for each collection but does not support an integrated search on the whole web archive. In other words, the search interface shown in Figure 1 is provided for each collection in WAS, and there is no such interface for the whole web archive of WAS. IA's Wayback Machine archive is accessible through the integrated search portal for all the IA collections. The UK Government Web Archive is integrated with the live government websites. When a requested URL is no longer available from the UK government live website, the request is automatically redirected to the UK Government Web Archive.

Unsupported search parameters: None of the web archives support searching for identical copies using MD5 fingerprint, or searching based on genre and update frequency. Safe search to filter out adult content was not found in any of the ten web archives. The scope and highly selective nature of some of these web archives may make such a safe search feature unnecessary. For example, the Government of Canada Web Archive and UK Government Web Archive only archive government websites; it is thus highly unlikely that those archives include adult content. Web archives that do bulk harvesting, such as the IA Wayback Machine, are likely to include adult content. The author was able to find adult websites using the IA Wayback Machine. For example, according to the search results page of IA, the porn website http://www.pornhub.com/ has been crawled 429 times since February 1, 2001. There are no access restrictions on these archived versions.

Usability issues on the search interface: Although not intended, some usability issues affecting functionality were identified. WAX provides a very simple Google-like search box, and there is no advanced search option on the search interface. The "help" link, which provides instructions for conducting advanced searches, is at the bottom of the page, far away from the search box itself, as shown in Figure 3.



HARVARD UNIVERSITY LIBRARY
Web Archive Collection Service

Harvard's Web Archive Collection Service (WAX) is part of the University Library's central infrastructure for the capture, management, storage, and display of web sites for long-term archiving. More About WAX

BROWSE COLLECTIONS

SL Sites: Archived Websites from Schlesinger Library Collections
Arthur and Elizabeth Schlesinger Library on the History of Women in America, Radcliffe Institute for Advanced Study

Capturing Women's Voices on the Web
Arthur and Elizabeth Schlesinger Library on the History of Women in America, Radcliffe Institute for Advanced Study

Constitutional Revision in Japan Research Project/憲法改正論議に関する研究
Edwin O. Reischauer Institute of Japanese Studies, Harvard University

A-Sites: Archived Harvard Web Sites
Harvard University Archives, Harvard University Library

SEARCH COLLECTIONS

Full text search of archived web sites:

☑ SL Sites: Archived Websites from Schlesinger Library Collections

☑ Capturing Women's Voices on the Web

☑ Constitutional Revision in Japan Research Project/憲法改正論議に関する研究

☑ A-Sites: Archived Harvard Web Sites

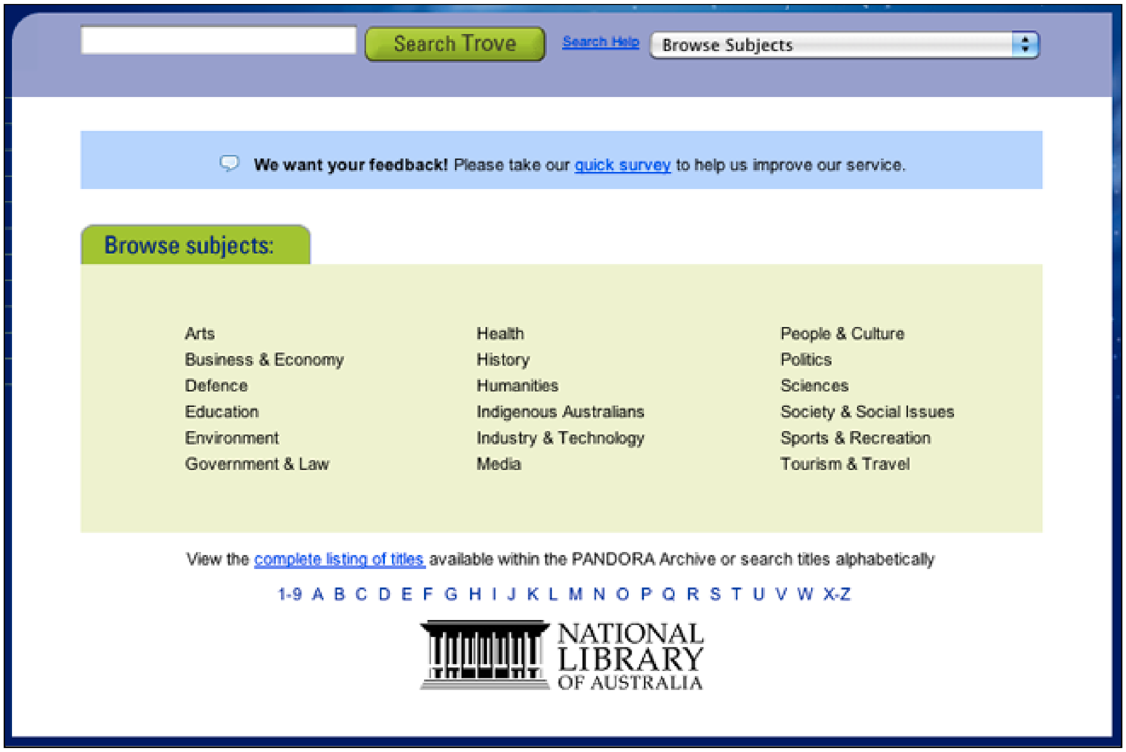
Use of WAX is subject to our Terms of Use
Harvard University Library Web Archive Collection Service
About WAX | Questions and Comments | Help

Harvard Home | HOLLIS Catalog | Harvard Libraries Portal
© 2008 President and Fellows of Harvard College

Figure 3: The search interface of WAX.

The instructions on the help page of WAX require users to use a certain syntax to conduct advanced searches. For example, to limit a search to files of a particular PDF format, the user needs to type in the search box: "type: application/pdf". In this example, users also need to memorize and type in the MIME type instead of choosing one from a list. At a time when search interfaces are becoming more user-centered and easier to use, this interface looks primitive. The very short history of this web archive may explain this obvious usability issue. However, similar problems also exist in web archives that have a longer history, although to a lesser degree. PANDORA and Library of Congress also provide only one simple search interface, and leave instructions for advanced search only in the "help" text. See Figure 4.

The PANDORA "help" page also provides a link to the advanced search interface. This link should be provided on the same web page as the single search box, since according to Ras & Bussel (2007), few users will use the "help" page.



We want your feedback! Please take our quick survey to help us improve our service.

Browse subjects:

Arts
Business & Economy
Defence
Education
Environment
Government & Law
Health
History
Humanities
Indigenous Australians
Industry & Technology
Media
People & Culture
Politics
Sciences
Society & Social Issues
Sports & Recreation
Tourism & Travel

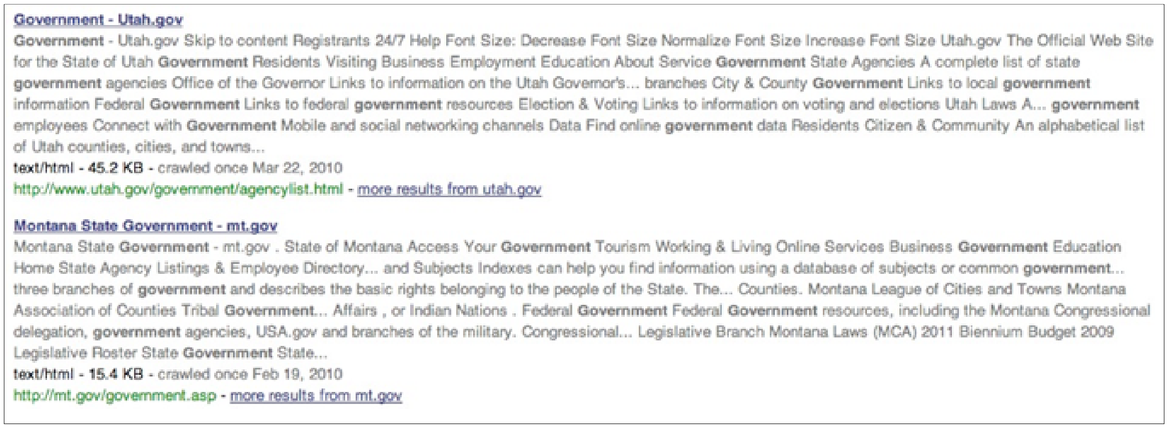
View the complete listing of titles available within the PANDORA Archive or search titles alphabetically

1-9 A B C D E F G H I J K L M N O P Q R S T U V W X-Z

Figure 4: The search and browse interface of PANDORA.

Search results

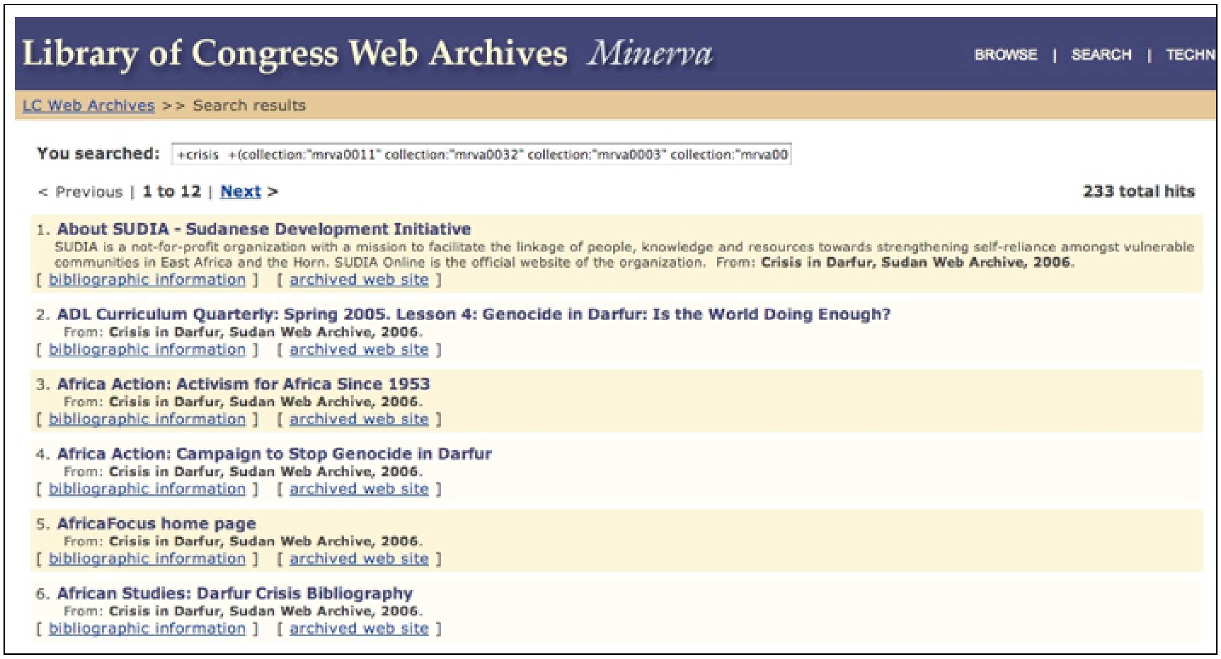
All the web archives present all archived versions of a URL, the date and time when each version was captured, and group all archived versions of a URL together. Five web archives provide a content summary for each returned URL that is automatically extracted from the archived web page. See Figure 5.


Government - Utah.gov
Government - Utah.gov Skip to content Registrants 24/7 Help Font Size: Decrease Font Size Normalize Font Size Increase Font Size Utah.gov The Official Web Site for the State of Utah **Government** Residents Visiting Business Employment Education About Service **Government** State Agencies A complete list of state **government** agencies Office of the Governor Links to information on the Utah Governor's... branches City & County **Government** Links to local **government** information Federal **Government** Links to federal **government** resources Election & Voting Links to information on voting and elections Utah Laws A... **government** employees Connect with **Government** Mobile and social networking channels Data Find online **government** data Residents Citizen & Community An alphabetical list of Utah counties, cities, and towns...
text/html - 45.2 KB - crawled once Mar 22, 2010
http://www.utah.gov/government/agencylist.html - more results from utah.gov

Montana State Government - mt.gov
Montana State **Government** - mt.gov . State of Montana Access Your **Government** Tourism Working & Living Online Services Business **Government** Education Home State Agency Listings & Employee Directory... and Subjects Indexes can help you find information using a database of subjects or common **government**... three branches of **government** and describes the basic rights belonging to the people of the State. The... Counties. Montana League of Cities and Towns Montana Association of Counties Tribal **Government**... Affairs , or Indian Nations . Federal **Government** Federal **Government** resources, including the Montana Congressional delegation, **government** agencies, USA.gov and branches of the military. Congressional... Legislative Branch Montana Laws (MCA) 2011 Biennium Budget 2009 Legislative Roster State **Government** State...
text/html - 15.4 KB - crawled once Feb 19, 2010
http://mt.gov/government.asp - more results from mt.gov

Figure 5: Content summary in the search results of Archive-It.

The Library of Congress and the New Zealand Web Archive provide a bibliographic record for each returned result instead of a content summary. See Figure 6.



Library of Congress Web Archives *Minerva*

BROWSE | SEARCH | TECHN

LC Web Archives >> Search results

You searched: +crisis +(collection:"mrva0011" collection:"mrva0032" collection:"mrva0003" collection:"mrva00

< Previous | 1 to 12 | Next >

233 total hits

1. **About SUDIA - Sudanese Development Initiative**
SUDIA is a not-for-profit organization with a mission to facilitate the linkage of people, knowledge and resources towards strengthening self-reliance amongst vulnerable communities in East Africa and the Horn. SUDIA Online is the official website of the organization. From: **Crisis in Darfur, Sudan Web Archive, 2006.**
[bibliographic information] [archived web site]

2. **ADL Curriculum Quarterly: Spring 2005. Lesson 4: Genocide in Darfur: Is the World Doing Enough?**
From: **Crisis in Darfur, Sudan Web Archive, 2006.**
[bibliographic information] [archived web site]

3. **Africa Action: Activism for Africa Since 1953**
From: **Crisis in Darfur, Sudan Web Archive, 2006.**
[bibliographic information] [archived web site]

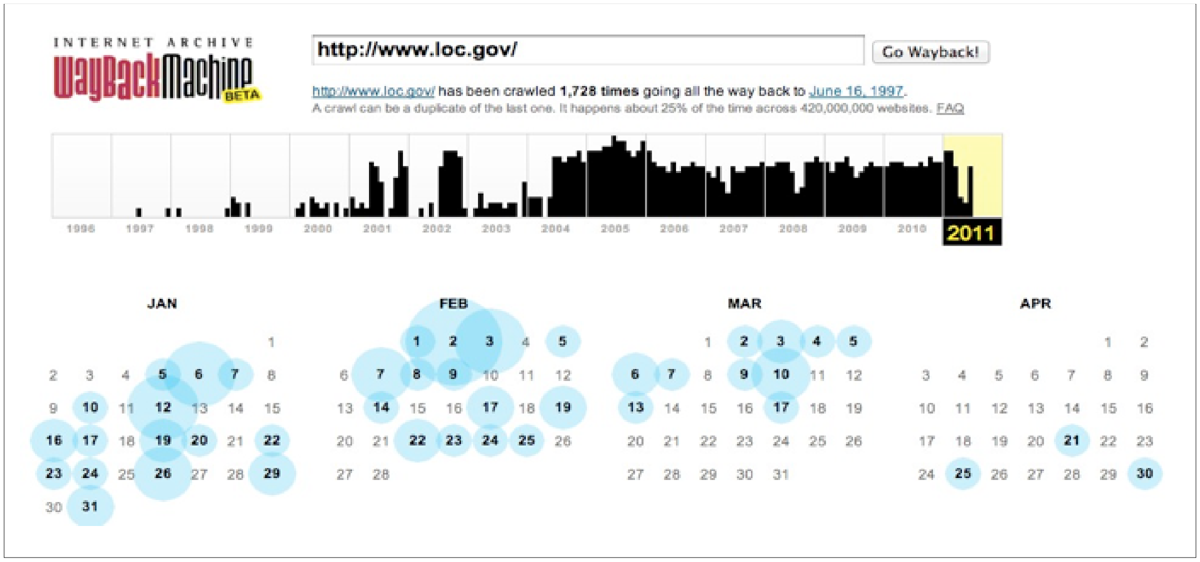
4. **Africa Action: Campaign to Stop Genocide in Darfur**
From: **Crisis in Darfur, Sudan Web Archive, 2006.**
[bibliographic information] [archived web site]

5. **AfricaFocus home page**
From: **Crisis in Darfur, Sudan Web Archive, 2006.**
[bibliographic information] [archived web site]

6. **African Studies: Darfur Crisis Bibliography**
From: **Crisis in Darfur, Sudan Web Archive, 2006.**
[bibliographic information] [archived web site]

Figure 6: Search results page from the Library of Congress Web Archives.

Other web archives return only matching URLs with their capture dates. IA Wayback Machine presents the resulting URLs in the format of a calendar. See Figure 7. The UK Web Archive presents the resulting URLs as a list as shown previously in Figure 2. All the web archives support navigation through hyperlinks, just like surfing on the Web.



http://www.loc.gov/ has been crawled 1,728 times going all the way back to June 16, 1997.
A crawl can be a duplicate of the last one. It happens about 25% of the time across 420,000,000 websites. FAQ

Figure 7: The calendar view of search results at IA Wayback Machine.

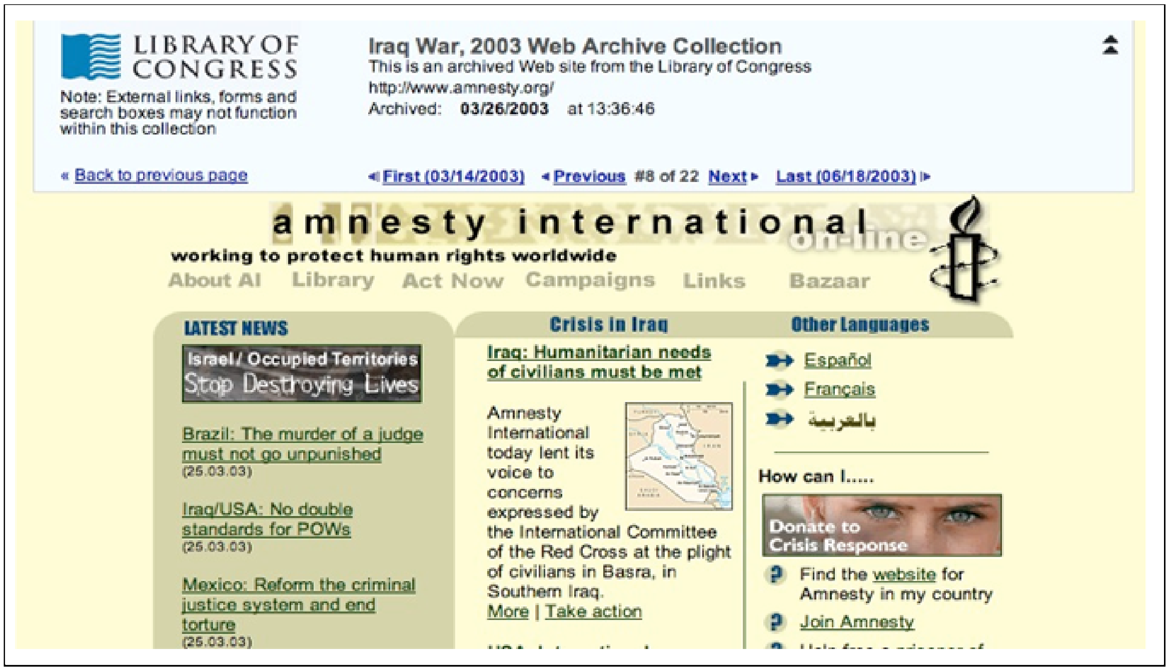
Group search results by domain: Five web archives show the hierarchical relationships between websites and the pages that belong to those websites. In other words, all of the web pages from the same web site that match the search criteria are grouped together in the search results. (See the example from Archive-It in Figure 5.) The UK Web Archive and UK Government Web Archive return matching web pages from the same website as different search results. Fortunately, pages from the same website are listed adjacent to each other. Thus, users are still able to see all web pages from the same site, although the results are not hierarchically presented. (See the earlier example from the UK Web Archive in Figure 2.)

Persistent identifiers: Four web archives provide persistent identifiers for archived web pages. PANDORA provides persistent identifiers for citing a title (a journal, a website, etc.) and the component parts of a title (e.g. an article within an issue of an electronic journal, or an image or a table within a web site). The persistent identifier for a title is ready-made. The persistent identifiers for a part of a title can be automatically generated upon request by using the citation service. The Library of Congress Web Archives provides a citation ID in its metadata records for websites and suggests that users include this citation ID when citing the archived web page. WAX offers ready-made citations in three styles — APA, Chicago and MLA — for three levels of resources: a web archive collection, a website and an individual archived version of a website. A URL is included in the ready-made citation format. Since this URL is used for the citation, it should be a persistent identifier. WAS displays persistent identifiers (archival URL) in the banner on top of the browser window for each archived web page.

Indicating archived content: All web archives construct URLs for archived web content based on the URL of the web archive and the original URL of the archived web page. For example, by looking at this URL:

> http://webarchive.loc.gov/lcwa0003/20030326083646/http://www.amnesty.org/

a user with the right experience knows that it is an archived web page from www.amnesty.org, preserved by the Library of Congress Web Archives, and captured at 8:36:46 am on March 26, 2003. However, this URL in the browser address bar is neither noticeable nor understandable to users who do not know how the URLs are constructed. A better approach to indicating archived web pages is to display a banner at the top of the archived page. All web archives except PANDORA use a banner to indicate that the version being viewed is an archived web page. IA Wayback Machine and Archive-It provide the option to hide or close the banner. While the option to close the banner is useful, users who later need the information in the banner cannot re-open it. The Library of Congress Web Archives and the UK Web Archive provide better designs. They allow users to hide the banner partially and recover the whole banner when needed. See Figure 8.



LIBRARY OF CONGRESS

Note: External links, forms and search boxes may not function within this collection

Iraq War, 2003 Web Archive Collection

This is an archived Web site from the Library of Congress

http://www.amnesty.org/

Archived: **03/26/2003** at 13:36:46

« Back to previous page

◂First (03/14/2003) ◂Previous #8 of 22 Next▸ Last (06/18/2003)▸

Brazil: The murder of a judge must not go unpunished
(25.03.03)

Iraq/USA: No double standards for POWs
(25.03.03)

Mexico: Reform the criminal justice system and end torture
(25.03.03)

Iraq: Humanitarian needs of civilians must be met

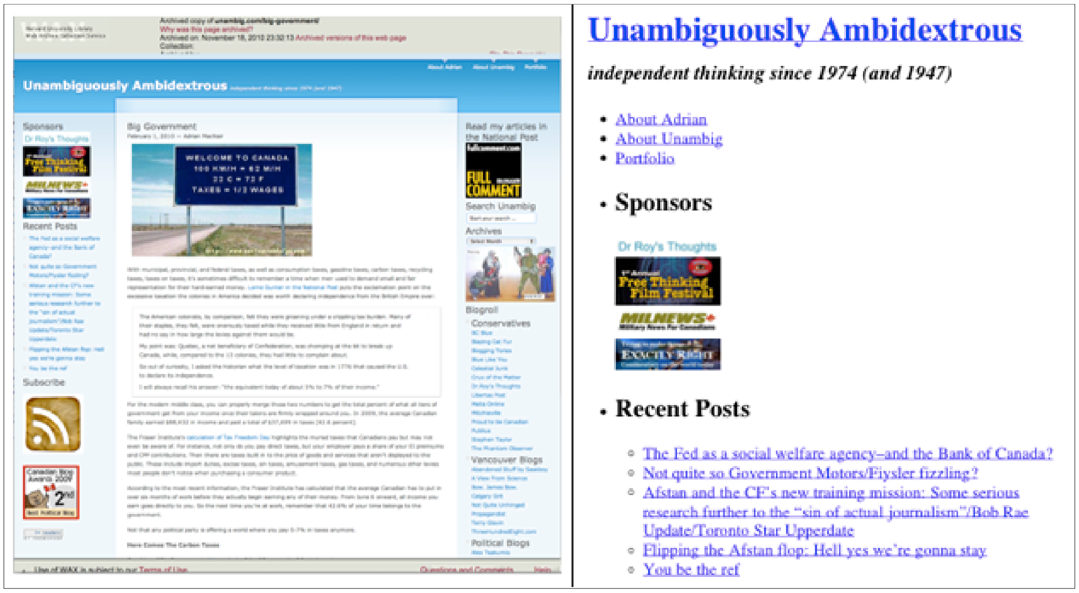
Amnesty International today lent its voice to concerns expressed by the International Committee of the Red Cross at the plight of civilians in Basra, in Southern Iraq.
More | Take action

Find the website for Amnesty in my country

Join Amnesty

Figure 8: The banner from the Library of Congress Web Archives. The double arrow at the top right corner of the banner allows users to partially hide the banner.

Printing: Three web archives do not allow users to print the banner at the top of the browser window. Two web archives restrict certain images from printing, and the layout of the prints is also different from what is presented in the browser. Figure 9 compares the two layouts. The difference between the two is so dramatic that without careful examination a user may not recognize that the two printed pages are actually the same web page.


Archived copy of unambig.com/big-government
Why was this page archived?
Archived on: November 18, 2010 23:32:13 Archived versions of this web page
Collection:
Unambiguously Ambidextrous
Sponsors
Dr Roy's Thoughts
Recent Posts
Subscribe
Big Government
WELCOME TO CANADA
100 KM/H = 62 M/H
22 C = 72 F
TAXES = 1/2 WAGES
Read my articles in the National Post
FULL COMMENT
Search Unambig
Archives
Blogroll
Conservatives
Vancouver Blogs
Political Blogs
Use of WAX is subject to our Terms of Use.
Questions and Comments
Help

Unambiguously Ambidextrous

independent thinking since 1974 (and 1947)

- About Adrian
- About Unambig
- Portfolio
- **Sponsors**

  Dr Roy's Thoughts
- **Recent Posts**
  - The Fed as a social welfare agency–and the Bank of Canada?
  - Not quite so Government Motors/Fiysler fizzling?
  - Afstan and the CF's new training mission: Some serious research further to the "sin of actual journalism"/Bob Rae Update/Toronto Star Upperdate
  - Flipping the Afstan flop: Hell yes we're gonna stay
  - You be the ref

Figure 9: Left image: the view of an archived page in the browser from WAX. Right image: the print view of the same page.

Certification of reliability: None of the web archives mention that they certify the reliability of records archived in their web archives. On the contrary, several web archives (IA Wayback Machine and WAX) provide a disclaimer saying that they are not responsible for the accuracy or reliability of content in their web archives. PANDORA says it takes various measures to preserve the authenticity and integrity of archived resources, but does not mention that the archive can certify the authenticity, integrity and reliability of the archived web content. The Government of Canada Web Archive preserves government websites, which are more likely to be used for government accountability. However, even that web archive provided a disclaimer for the reliability and authenticity of its archived web resources.

Three items on the functionality checklist are about the comparison of different versions of an archived URL. IA Wayback Machine is the only web archive that provides the functionality of comparing two versions of a web page. However, this functionality did not work during the time this study was conducted, so it is not known to what extent, and how, IA Wayback Machine supports the comparison of different versions of web pages.

Browsing

Eight of the ten web archives provide browsing functionality. Seven of them organize content into collections or categories based on subject or genre, as shown in Figure 4. The Government of Canada Web Archive organizes websites by government departments. The collection of IA Wayback Machine is very broad and there is no human generated metadata that helps categorize its collection. However, in the future, its collection might be organized into categories and sub-categories based on automatically extracted metadata. For example, the collection could be categorized first by country domains, then the web of a country could be categorized by other Top Level Domains (.gov, .edu, etc.), each sub-category could be further divided by genre (blogs, newspapers, virtual worlds), then categorized by media types (PDF, HTML, video, etc.), and so on. This browsing structure would allow users to use the web archive when they do not know a specific URL.

Policy-related functionalities

Policies for handling users' requests to block public access to content already in an archive, to remove content from an archive, or to add content to an archive, are not consistent across the ten archives. For example, blocking public access is only mentioned explicitly by WAS; the Library of Congress proactively blocks online access to websites that they do not have permission from copyright owners to show; and PANDORA proactively restricts access to a very small portion of titles for commercial reasons, or because their content is sensitive.

Regarding users' requests to add or remove specific URLs, WAS and six other web archives state that they have a "takedown policy" for removing specific web pages from the archive upon users' requests. The Library of Congress explicitly states that it does not accept user requests to add specific URLs to their archive. The UK Government Web Archive and WAX do not mention whether they allow users to request specific URLs to be archived, but some of the other web archives do accept user-recommended URLs for archiving.

Personalized services

The National Library of New Zealand catalog provides personalized service for users, such as saving searches and records for later user. Since the New Zealand Web Archive is accessible through the library catalog, the personalized service is also available for the web archive users. However, it was not created specifically for the web archive. IA Wayback Machine allows users to create an account, but this personal account is used for uploading and sharing information to other databases of IA Wayback Machine and seems unrelated to the Wayback archive. No other web archive was found to provide personalized services. PANDORA is the only web archive that publishes monthly reports about new titles added to the web archive. This public notification is not the same as the personalized alert service mentioned in the IIPC use cases. Personalized services are a common functionality in many traditional library OPACs and digital libraries. They are also provided by some online archives with blended web resources and digitized content, such as the UCLA Online Campaign Literature Archive of the University of California, Los Angeles. By adding this functionality, web archives are more likely to be capable of attracting and keeping repeat users.

Data mining

None of the web archives show or mention that they support any of the data mining functionalities on the checklist. The UK Web Archive offers two visualization interfaces based on the mining of its archived content: tag clouds and a 3D wall. However, it does not seem to provide data mining functionalities for researchers. None of the web archives were found to preserve website log files that include information about operating systems, web servers, versions, and network connection speeds. Thus, these web archives are unlikely to be able to support the kind of research about the evolution of web technology that is mentioned in IIPC use cases, although the archived web pages themselves demonstrate the evolution of web technology. Website log files are more likely to be preserved for an organization's internal records management purposes or for specific research purposes. All of the web archives in this study rely mainly on online harvesting for external archiving, and this at least partially explains why they do not archive log files.

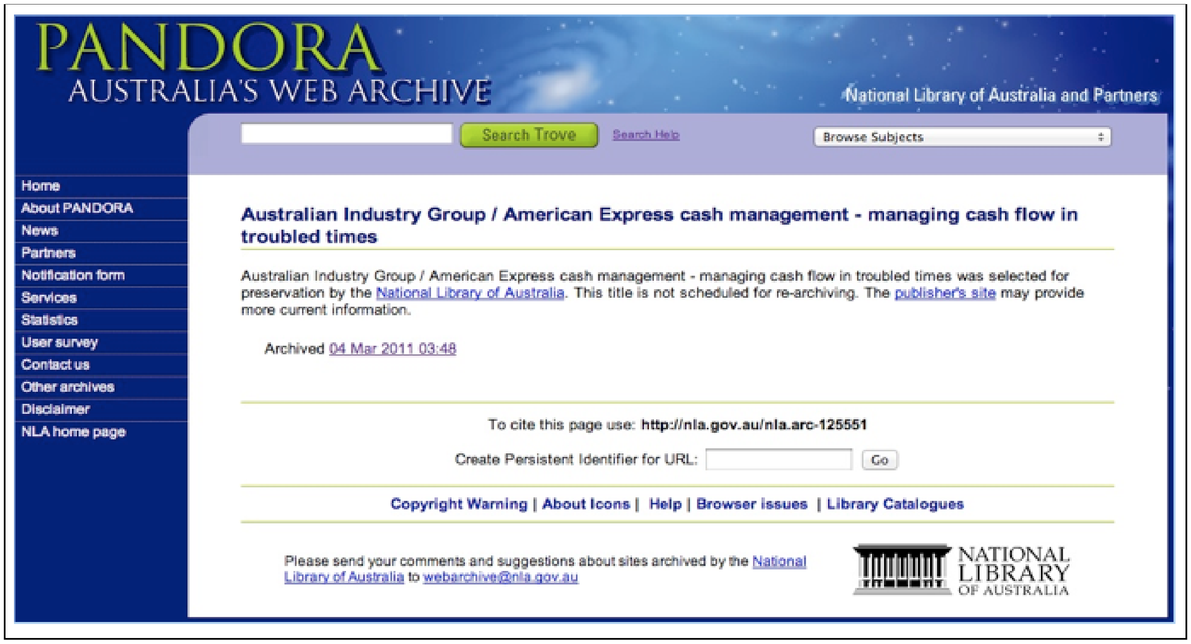
Reconstruction of lost websites

Web archives have the potential to recover at least portions of lost websites. The tool WARRICK created by Frank McCown for recovering lost websites relies on data from IA Wayback Machine and the caches of Google, MSN, and Yahoo (Internet Archive Web Team, 2007). Nevertheless, none of the ten web archives evaluated in this study mention that they support the recovery of lost websites. Two web archives mentioned there is a delay between the crawling date and the date when the crawled resources are accessible through the web archive. The delay in WAX is at least three months (Harvard University Library, 2009). In the case of IA Wayback Machine, the delay is 6-12 months (Archive-It, 2011). This causes problems for website owners who want to recover a recently lost website as soon as possible.

Functionalities discovered in the study that are not on the checklist

Duplicates management: IA Wayback Machine provides the option of showing duplicates in search results.

Discoverability by search engines: Several web archives explicitly state that they block search engine crawlers from indexing archived web pages. The Government of Canada Web Archive only allows Google to index its main page. The UK Web Archive and PANDORA allow commercial search engines to index the information page of each archived title. Titles archived by the UK Web Archive and PANDORA will show up in Google search results if the right search query is used. For example, the query "Australian Industry Group/American Express cash management — managing cash flow in troubled times" will return the archived version at PANDORA. The search result will not lead users directly to the archived page. Instead, it brings the user to the information page about this archived document on the PANDORA website. See Figure 10. The web page clearly shows the requested document is archived by PANDORA and keeps users from confusing it with the live website. A similar mechanism is used by the UK Web Archive.



Australian Industry Group / American Express cash management - managing cash flow in troubled times

Australian Industry Group / American Express cash management - managing cash flow in troubled times was selected for preservation by the National Library of Australia. This title is not scheduled for re-archiving. The publisher's site may provide more current information.

Archived 04 Mar 2011 03:48

To cite this page use: http://nla.gov.au/nla.arc-125551

Create Persistent Identifier for URL:

Please send your comments and suggestions about sites archived by the National Library of Australia to webarchive@nla.gov.au

Figure 10: A Google search result for a web page archived by PANDORA.

Indicating non-archived content: When a requested URL is not archived, all of the web archives return a page showing the requested URL, why the page is not archived, and/or what users should do. IA Wayback Machine also automatically captures the requested URL, and notifies users about this capture in the banner on top of browser window.

Conclusions

This paper analyzed the functionalities of ten established web archives in April 2011. Some basic functionalities, such as searching by URL and keywords, and narrowing the search by dates, domain and media type, are commonly supported by the web archives. However, more advanced functionalities, such as data mining, personalized services, and the reconstruction of lost websites, are not supported by any of the investigated web archives. Several web archives were found to have obvious usability issues, such as invisible help links, hidden advanced search tools, and segmented searching.

These findings also show where the web archives are in terms of functionalities and how they compare to the IIPC use cases and identified user expectations. The existence of the problems does not necessarily mean that the web archives do not care or do not have the ability to solve these issues. It is more likely that they currently focus their attention on collection building and do not have the time to provide more advanced functionalities and improve the usability of their web archives. This is especially true for some of the newer web archives.

The author believes that the archives will improve their functionalities and usability over time. In fact, when the author revisited some of the archives in August 2011, approximately four months after the initial study, some improvements were noted. For example, the UK Web Archive had added a *n*-gram search functionality, which can display a graph showing how the search term has occurred in the UK Web Archive over time. The UK Government Web Archive had also added a functionality that automatically clusters the search results by subjects, and allows users to filter the search results by subjects.

Limitation of the study

This study relied solely on data gathered from the websites of ten publicly accessible web archives with an English interface. This poses concerns about the completeness of the findings. It is possible that some web archives that are accessible on-site only, and non-English web archives, provide more advanced functionalities that were not discovered in this study. It is also possible that some functionalities, such as extracting a subset of the web archive for outside analysis and recovering portions of lost websites, are supported but not visible on the websites of the web archives. A more thorough study should include gathering information from the people who build and maintain the web archives, in addition to examining the archives themselves.

Note

This study followed a comprehensive literature review and evaluation of web archiving methods conducted by this author and reported in the article, "An Overview of Web Archiving", also published in the March/April 2012 issue of *D-Lib Magazine*.

References

[1] Archive-it. (2011). FAQ. *Archive-it*.

[2] Costa, M. & Silva M. J. (September 2010). Understanding the information needs of Web archive users. *10th International Web Archiving Workshop*, Vienna.

[3] Harvard University Library. (2009). WAX Public Interface Help.

[4] International Internet Preservation Consortium Access Working Group. (2006). Use cases for access to Internet Archives. *International Internet Preservation Consortium*.

[5] Internet Archive Web Team. (2007, February 22). Warrick, a tool for recovering websites. Message posted.

[6] Jatowt, A., Kawai, Y. & Tanaka, K. (September, 2008). Using page histories for improving browsing the Web. *8th International Workshop for Web Archiving*, Denmark.

[7] Niu, J. (2012). An Overview of Web Archiving. *D-Lib Magazine*, Vol.18, No. 3/4. http://dx.doi.org/10.1045/march2012-niu1.

[8] Ras, M. & Bussel, S. V. (July 2007). Web archiving user survey.

[9] Tahmasebi, N., Zenz, G. & Iofciu, T. (September, 2010). Terminology Evolution Module for Web Archives in the LiWA Context. *10th International Web Archiving Workshop*, Vienna.

Appendix 1: Functionality test instrument for web archives

Basic information about the web archive:

1. Name:
2. URL:
3. Creator:

Which of the following functions are supported by the web archive? Choose all that apply:

Search parameters:

- Search by URL
- Search by keyword
- Boolean search
- Narrow down search results by date
- Search within one specific archived website
- Search for identical copies using MD5 fingerprint
- Search by web content characteristics rather than by keywords (e.g. update frequency, genre, media type)
- The web archive is accessible through a federated search portal
- Users can get access to log files of archived websites that store http response headers, download times, transfer speeds and other relevant information.
- Safe search to filter adult contents
- Distinction between simple searching and advanced searching

Search results:

- Search results are ranked based on relevancy to the query
- Search results are grouped by domain
- Users can see all archived versions of a URL
- Users can see the dates on which each URL is archived
- Each returned result has a content summary
- Users are informed if an archived webpage has changed during a period of time
- Users are informed about what parts were changed on a webpage between different archival dates
- Users can get a summary of the differences of two version of an archived web page (e.g. percentage, word count, link count, image count)
- Users can navigate the archived website as it was on the internet
- Each archived web page has a persistent identifier to refer to
- The web archive allows users to print out relevant pages
- The web archive provides signed declaration of the reliability of printouts
- Indicating archived website
- Archived websites are the same as the original version of this site
- Indicate how a website can be accessed (online or onsite)

Browsing:

- Collections in the web archive are organized into a hierarchy; users can browse the hierarchy of the collection, narrow down by time period and domain, and browse to the intended result.

Policy-related functionalities:

- Users can request certain URLs to be archived.
- Public access to certain web content can be blocked upon owner's request (take down)

Personalized services:

- The web archive provides personalized web archive user interface (allow personal account). For example, it allows users to save performed searches for reference and possible re-searching, to bookmark some web pages into a personal bookmark folder, and to save some documents in the web archives for further purposes.
- The web archive actively notifies users about updates in the web archive by email. For example, it notifies users every time harvests find a certain page has been updated or changed, specifying the specific changes that have been made to a URL during a period of time.
- The user will be notified through email when the URL is harvested and when the webpage becomes available through the web archive.

Data mining:

- The web archive can return graphs illustrating how certain archived websites associate with certain events in a period of time.

- The web archive can provide linking information for an archived web page (incoming links, outgoing links, internal links).
- The web archive allows users to extract a subset of the archive based on specific criteria (time-span, domain name, language, file format, metadata either alone or in combination), and exports the extracted subset for processing elsewhere.
- The web archive allows users to extract a subset of the archive based on specific criteria (time-span, domain name, language, file format, metadata either alone or in combination), and then process and analyze data in the web archive.

Reconstruction of lost websites:

- Users can use the web archives to reconstruct a certain version of a lost website, retaining the structure of the original web site.
- Proven owners of a website can use the web archives to reconstruct a certain version of their lost website, retaining the structure of the original web site.

About the Author

Jinfang Niu is an assistant professor at the School of Information, University of South Florida. She received her Ph.D. from the University of Michigan, Ann Arbor. Prior to that, she worked as a librarian at the Tsinghua University Library for three years. Her current research focuses on electronic records and digital curation.