Building Thematic Web Collections: Challenges and Experiences from the September 11 Web Archive and the Election 2002 Web Archive

Steven M. Schneider[1], Kirsten Foot[2] , Michele Kimpton[3] , Gina Jones[4]

[1] SUNY Institute of Technology & WebArchivist.org 12 North-Horatio Street
Utica, NY 13502
steve@sunyit.edu
[2] Department of Communication, University of Washington & WebArchivist.org, Seattle, WA 98101
kfoot@u.washington.edu
[3] Internet Archive, P.O. Box 29244 Presidio of San Francisco San Francisco, CA 94129-0244
michele@archive.org
[4] Library Services, Library of Congress 101 Independence Ave, SE Washington, DC 20540
gjon@loc.gov

Abstract. One method for creating large-scale collections of Web materials is to use a "thematic" approach. In this paper, we introduce the concept of a thematic Web collection, discuss the experience of our organizations which have collaborated in the development and presentation of two thematic Web collections, identify challenges associated with thematic archiving, and comment on the value of thematic archiving from a library, archivist and scholarly perspective.

1 Thematic Web Archiving

A thematic Web collection is an archive of Web objects identified and captured using a set of URLs believed to be relevant to a specific theme or topic. A set of carefully selected URLs is used as the "seeds" for collecting activity. These URLs, representing either sites or pages of interest, are crawled at an established periodicity, and with clearly specified rules concerning the crawling of linked pages and objects. For example, a crawler might be instructed to start at a given set of base URLs, to crawl all pages and page requisites with URLs from within the domain of the base URLs, and to repeat this crawling procedure once per week for six months.

Thematic collections can be contrasted to broad-based crawling activities. In broad-based crawling activities, a Web crawling program will start its archiving activity from a set of URLs that have no particular content relationship to each other. The Web crawling

program collects objects to be included in the archive by following links from the seed objects. This automated collection process continues indefinitely until all paths from the seed objects are exhausted or time and resources available reach their limit.

1.1 Examples of Thematic Web Archives Produced by the Library of Congress

This section of the paper examines two thematic Web archives produced by the United States Library of Congress. The September 11 Web Archive contains Web sites collected in the immediate aftermath of the September 11 terrorist attacks in the United States. The Election 2002 Web Archive includes Web sites related to the 2002 elections in the United States for U.S. Senator, U.S. Representative, and Governor of the various states in the U.S. and territories. Both Web archives were produced in collaboration with the Internet Archive and WebArchivist.org.

The U. S. Library of Congress is the largest library in the world, with collections that include more than 18 million books, 2.5 million recordings, 12 million photographs, 4.5 million maps, and 54 million manuscripts. The Library is the oldest federal cultural institution in the United States, and serves as the research arm of the U.S. Congress. Its mission is to make its resources available and useful to the Congress and the American people and to sustain and preserve a universal collection of knowledge and creativity for future generations.

The Internet Archive is a 501(c) 3 non-profit organization, and holds the largest collection of born-digital material in the world, totally over 150 terabytes of data. Its mission is to build a digital library of Internet sites and other cultural artifacts in digital form. These collections are freely available to researchers, historians, scholars, and the general public.

WebArchivist.org is a scholarly research group directed by professors Steve Schneider at the SUNY Institute of Technology and Kirsten Foot at the University of Washington. WebArchivist.org develops systems for identifying, collecting, cataloging and analyzing large-scale archives of Web objects, and supports scholarly and curatorial work associated with Web archives.

2 About the Collections: September 11, Election 2002

Both collections are comprised of Web sites collected as part of a joint effort between the Library's Recommending Officers and the WebArchivist.org. The table below provides a general outline of the acquisition plan and the construct of each archive.

Table 1. Collection Overview for the September 11 Web Archive and the Election 2002 Web Archive

	September 11 Web Archive	Election 2002 Web Archive
URLs	30,000+	3,000+
Crawl Dates	September 11, 2001 - December 1, 2001	July 1, 2002 – November 30, 2002
Crawler	Internet Archive/ALEXA crawler	Internet Archive/ALEXA crawler
Crawl Periodicity	Daily	Varied
Unique URLs	332,000	82,000
HTML/TEXT/RTF	193K (60%)	48K (59%)
IMAGES	127K (38%)	29K (35%)
Size of collection	5 TB	1 TB

2.1 Cataloging

Site-level descriptive metadata is in the process of being created and derived for the Election 2002 Web Archive Web sites and 2,500 Web sites from the September 11 Web Archive.

Metadata to be provided for selected sites include: (1) Name (Creator or issuing publisher); (2) Title; (3) Abstract; (4) Capture dates; (5) Genre; (6) Physical Description/Format; (7) Language; (8) Subject Heading, using a controlled vocabulary from simplified Library of Congress subject headings; and (9) Access Conditions or restrictions. Capture dates, title, and physical description are derived from a machine-based examination of index files and archived materials using Library of Congress specifications; the remaining data is created from a manual examination of archived materials using Library of Congress specifications, and manually entered into the record. The Library of Congress has contracted with WebArchivist.org to perform this cataloging work.

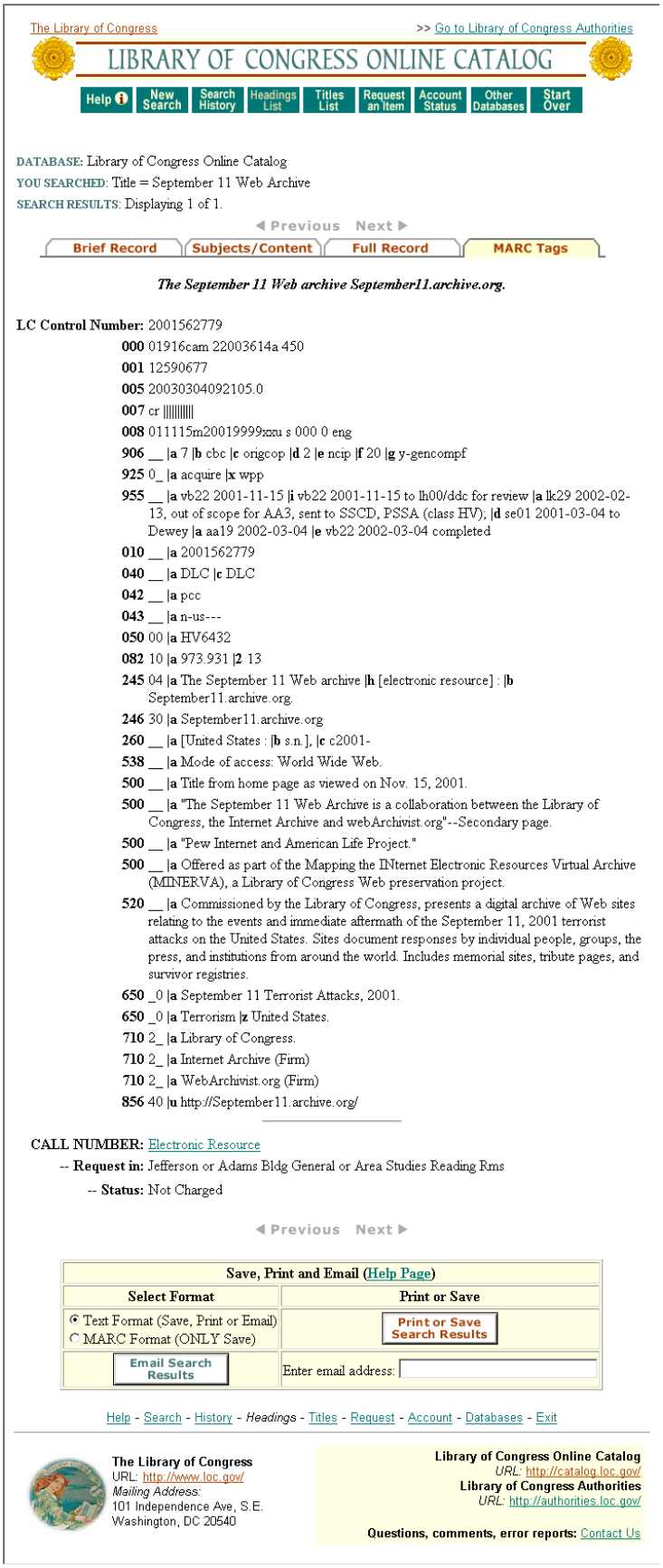
2.2 Management

Upon completion of the archiving process, Internet Archive processes and creates an Index for the collection. The collection and indices are copied and subsequently delivered on hard drives to the Library of Congress.

Upon receipt of the collection, the Library ports it to their servers and the collection is accessed using the Alexa Wayback program. System tools are used to analyze and characterize the collection. Site and page-level quality assurance checking is done to identify possible anomalies.

2.3 Access

The Library is working to provide access to these collections in a number of ways. All thematic Web archives will have a collection level record. Figure 1 provides a snapshot of the collection level bibliographic record for the September 11 Web Archive.



The Library of Congress >> Go to Library of Congress Authorities

LIBRARY OF CONGRESS ONLINE CATALOG

Help | New Search | Search History | Headings List | Titles List | Request an Item | Account Status | Other Databases | Start Over

DATABASE: Library of Congress Online Catalog
YOU SEARCHED: Title = September 11 Web Archive
SEARCH RESULTS: Displaying 1 of 1.

◄ Previous Next ►

Brief Record | Subjects/Content | Full Record | MARC Tags

The September 11 Web archive September11.archive.org.

LC Control Number: 2001562779

000 01916cam 22003614a 450
001 12590677
005 20030304092105.0
007 cr ||||||||||
008 011115m20019999xxu s 000 0 eng
906 __ |**a** 7 |**b** cbc |**c** origcop |**d** 2 |**e** ncip |**f** 20 |**g** y-gencompf
925 0_ |**a** acquire |**x** wpp
955 __ |**a** vb22 2001-11-15 |**i** vb22 2001-11-15 to lh00/ddc for review |**a** lk29 2002-02-13, out of scope for AA3, sent to SSCD, PSSA (class HV); |**d** se01 2001-03-04 to Dewey |**a** aa19 2002-03-04 |**e** vb22 2002-03-04 completed
010 __ |**a** 2001562779
040 __ |**a** DLC |**c** DLC
042 __ |**a** pcc
043 __ |**a** n-us---
050 00 |**a** HV6432
082 10 |**a** 973.931 |**2** 13
245 04 |**a** The September 11 Web archive |**h** [electronic resource] : |**b** September11.archive.org.
246 30 |**a** September11.archive.org
260 __ |**a** [United States : |**b** s.n.], |**c** c2001-
538 __ |**a** Mode of access: World Wide Web.
500 __ |**a** Title from home page as viewed on Nov. 15, 2001.
500 __ |**a** "The September 11 Web Archive is a collaboration between the Library of Congress, the Internet Archive and webArchivist.org"--Secondary page.
500 __ |**a** "Pew Internet and American Life Project."
500 __ |**a** Offered as part of the Mapping the INternet Electronic Resources Virtual Archive (MINERVA), a Library of Congress Web preservation project.
520 __ |**a** Commissioned by the Library of Congress, presents a digital archive of Web sites relating to the events and immediate aftermath of the September 11, 2001 terrorist attacks on the United States. Sites document responses by individual people, groups, the press, and institutions from around the world. Includes memorial sites, tribute pages, and survivor registries.
650 _0 |**a** September 11 Terrorist Attacks, 2001.
650 _0 |**a** Terrorism |**z** United States.
710 2_ |**a** Library of Congress.
710 2_ |**a** Internet Archive (Firm)
710 2_ |**a** WebArchivist.org (Firm)
856 40 |**u** http://September11.archive.org/

CALL NUMBER: Electronic Resource
-- **Request in:** Jefferson or Adams Bldg General or Area Studies Reading Rms
-- **Status:** Not Charged

◄ Previous Next ►

Save, Print and Email (Help Page)	
Select Format	**Print or Save**
◉ Text Format (Save, Print or Email) ○ MARC Format (ONLY Save)	Print or Save Search Results
Email Search Results	Enter email address:

Help - Search - History - *Headings* - Titles - Request - Account - Databases - Exit

The Library of Congress
URL: http://www.loc.gov/
Mailing Address:
101 Independence Ave, S.E.
Washington, DC 20540

Library of Congress Online Catalog
URL: http://catalog.loc.gov/
Library of Congress Authorities
URL: http://authorities.loc.gov/

Questions, comments, error reports: Contact Us

Figure 1. Snapshot of the collection level bibliographic record for the September 11 Web Archive

All of the Election 2002 Web Archive and 2,500 Web sites of the September 11 Web Archive will be cataloged using MODS (Metadata Object Description Schema), allowing a variety of access points through the cataloged metadata. Figure 2 demonstrates a catalog record for the Election 2002 Web Archive.

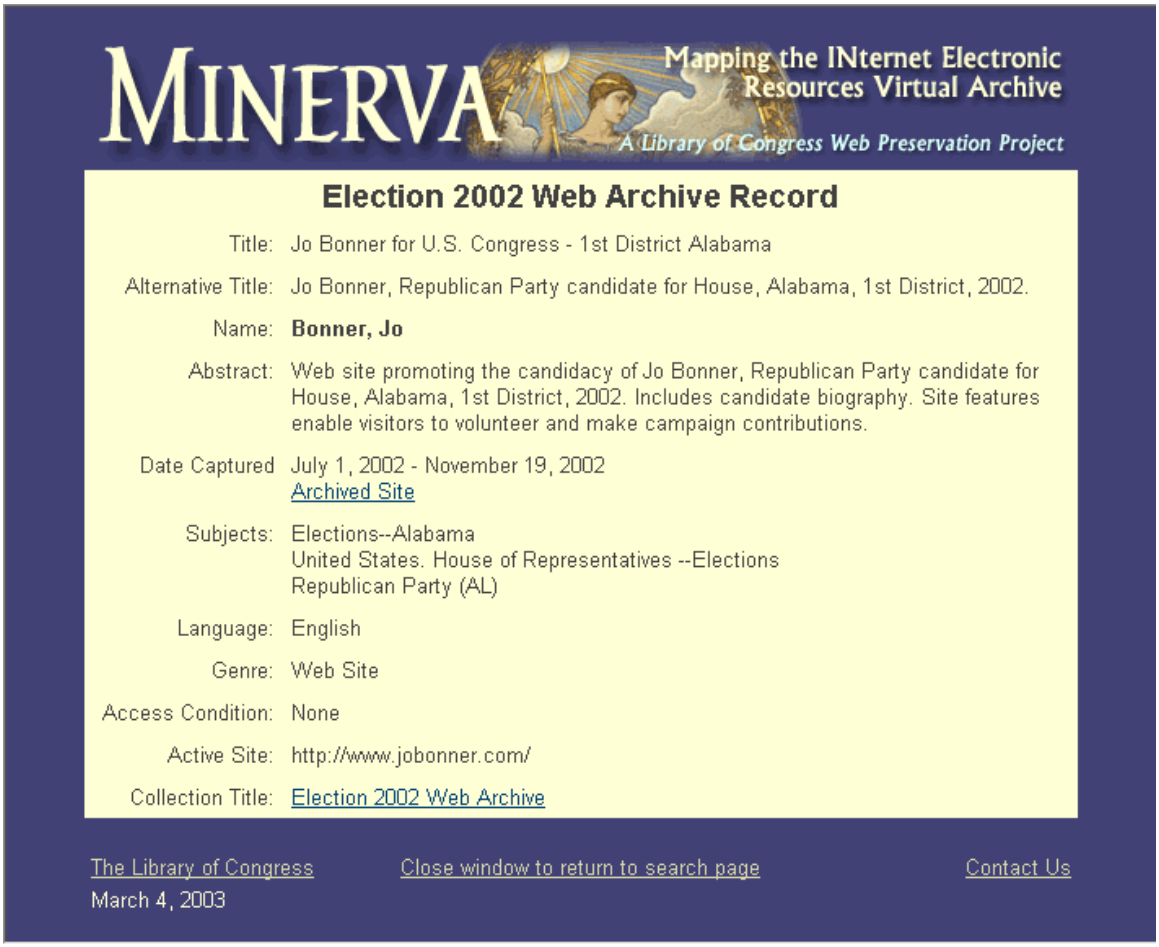


Figure 2. Snapshot of an Election 2002 Web Archive record

In addition to catalog entry points, the Library is experimenting with search engine indexing to allow for improved searchability of its collections. The challenge for the September 11 Web Archive is significant. Because less than 10% of approximately 30,000 Web sites will be cataloged, access to its over 331 million objects must be provided through some search interface. The Library of Congress is using its licensed Web page search engine (Inktomi®) to index the Web collections by indexing the initial target Web page captured (homepage or otherwise). Search terms entered into the search box will search the index, and the search results will provide access points, to both the indexed page that has those search terms and to the Web archive resource page that will allow access to all archived captures of that Web site via the Wayback program. Figure 3 provides an example of a search output results page using the search terms "pentagon", "navy", and "officers" in the September 11 Web Archive.

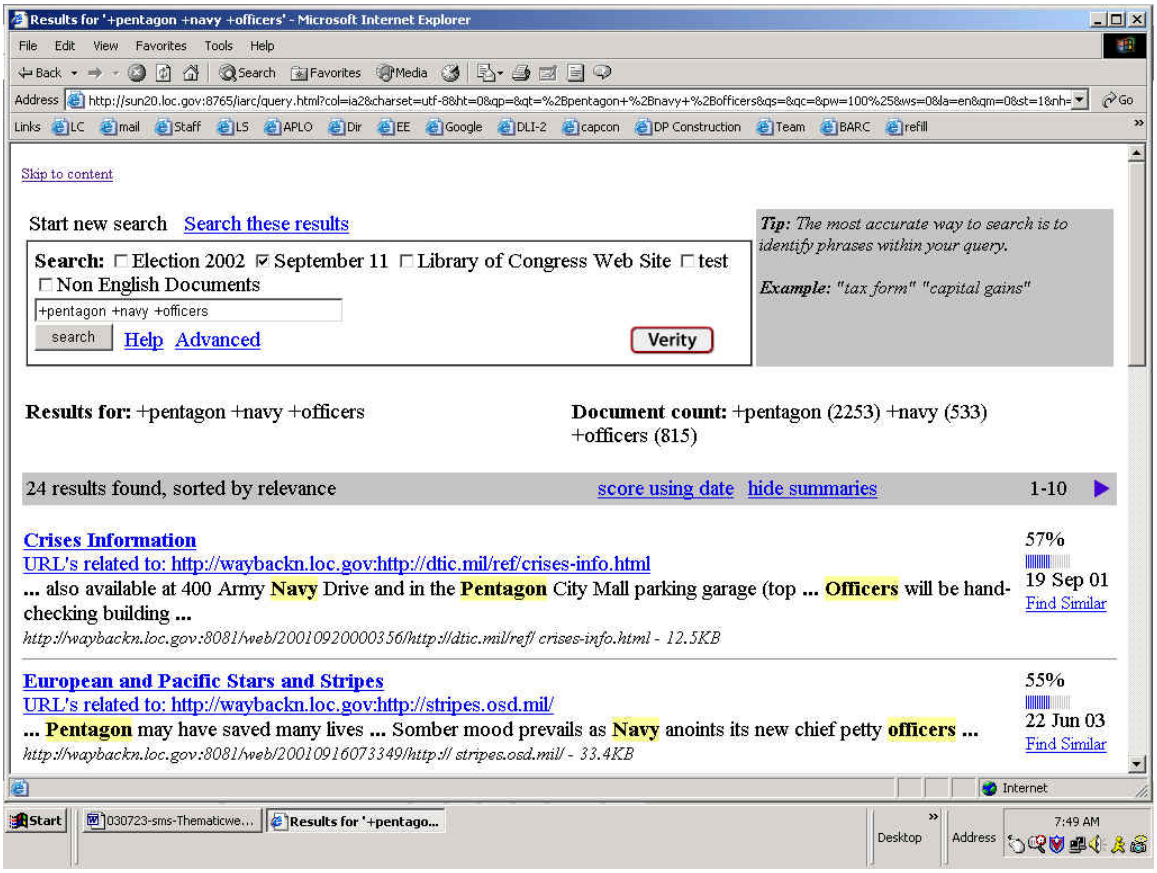


Figure 3. Examples of search term "+pentagon, +navy, +officers" in the September 11 Web Archive

3 September 11 Web Archive: Project Genesis and Current Status

Within days after the attacks on September 11, WebArchivist.org, the Internet Archive, and the Library of Congress had agreed to collaborate on a Web archiving collection. Previously, during the summer of 2001, Schneider and Foot, co-directors of the WebArchivist.org research group based at the SUNY Institute of Technology and the University of Washington, had contacted staff associated with the Library's MINERVA project to propose a collaboration Web archiving project around the 2002 federal elections. A structure for collaboration including complementary roles and sets of responsibilities had been sketched out by early September; this structure immediately formed the basis of the September 11 collaboration. The Library commissioned the Internet Archive to begin collecting daily impressions of URLs related to the attacks and their aftermath as they were identified. Library staff, WebArchivist.org researchers, and Web users around the world submitted URLs to be added to the collection; WebArchivist.org made accessible a browser-based tool that anyone could use to add URLs to the crawl list, and developed a method of submitting seed URLs to the Internet Archive for inclusion in the archiving process. The Pew Internet & American Life Project, www.pewinternet.org, provided funding to support the initial WebArchivist.org activities, and later analysis of Web sites collected in the archive.

The Library's Web archiving team that supported the September 11 Web Archive consisted of one reference librarian who served as the team leader, one cataloger, and personnel from Information Technology Services, all on a part-time status. The Library's Recommending Officers and staff from the National Digital Library selected sites for inclusion in the September 11 Web Archive. The Internet Archive was the collection agent for the September 11 Web Archive, and initiated and managed all crawling activity associated with the project. The Internet Archive received seed URLs from the Library of Congress and WebArchivist.org, and initiated daily crawls using those seeds. The Internet Archive provided initial public access to the collection, starting on October 11, 2001.

WebArchivist.org collaborated with the Library to identify sites for inclusion in the archive, and developed and implemented the strategy to catalog 2,500 sites in the collection in accordance with standards established by the Library. WebArchivist.org developed and hosted September11.archive.org, an interface to the collection, beginning on October 11, 2001. In addition, in collaboration with the Pew Internet & American Life Project, WebArchivist.org completed several scholarly analyses examining the nature of the Web in the post-September 11environment.

The Pew Internet & American Life Project provided initial funding to WebArchivist.org to support site identification activities, interface development, and scholarly analysis of the post-September 11 Web sphere. These analyses enhanced the user interface for the collection by generating producer type and action-based search fields, and resulted in a set of reports and exhibits, released in September 2002.

The Library of Congress' September 11 Web Archive is being processed for cataloging and access. When that process is complete, all Web sites within this collection will be available for use. In addition to the search function, full Web site/URL lists will be provided from the Library's Web site to provide access to the collection. Webarchivist.org is presently working with the Library to create a searchable metadata database for 2,500 Web sites. Sites selected for cataloging were identified through a relevance ranking that will hopefully provide rich September 11 Web content and critical Web expression.

4 Election 2002 Web Archive: Project Genesis and Current Status

Through a collaboration initiated by WebArchivist.org, the Election 2002 Web Archive project drew on lessons learned by the Library and the Internet Archive in producing the Election 2000: An Internet Library, and the experiences of Schneider and Foot in a parallel archiving and analysis project at the University of Pennsylvania around the 2000 elections, as well as the September 11 Web Archive. A grant from the Pew Charitable Trusts to the University of Washington enabled WebArchivist.org to work with the Library in developing a metadata-driven interface for the thematic Web archive that the Library engaged Internet Archive to collect. The collection plan for the Election 2002 Web Archive focuses on candidate, party, government, press, advocacy group and citizen sites related to the 2002 federal, gubernatorial collections and mayoral campaigns.

During the Election 2002 Web crawling and archiving process, the Library of Congress added personnel to its Web archiving team. A new full-time team leader, a technical specialist and a technical assistant joined the existing team to assist the efforts. The expanded team developed expertise during this project phase to assist in conducting quality assurance reviews of crawled Web sites. Additional cataloging specialists provided assistance and expertise for the cataloging process. Legal staff became permanent advisors to help develop cooperative agreements and to assist the team with rights management.

The Internet Archive served as the collection agent for the Library of Congress. The Library of Congress and WebArchivist.org selected web sites. Seed URLs were developed by WebArchivist.org and sent to the Internet Archive. The seed URLs were

crawled at the specified periodicity and with the specified depth levels. In addition, Internet Archive provided archive access within 24 hours of collection to staff affiliated with the Library of Congress and WebArchivist.org, facilitating verification, collection and analysis. Finally, Internet Archive provided public access to the collection, via the Library of Congress's Web interface, beginning in March 2003.

WebArchivist.org developed and implemented tools to facilitate the dynamic identification of Web sites for inclusion in the Election 2002 Web Archive, and systems to inform the collection agent of current seed URLs and their associated internal depth, external depth and collection periodicity. In addition, WebArchivist.org developed tools to facilitate the collection of metadata associated with the sites in the collection, and created an interface allowing users of the archive to identify sites of interest using this metadata.

The collection of Web materials for the Election 2002 Web Archive was fully funded by the Library of Congress. The Pew Charitable Trusts provided a substantial grant to the University of Washington for the project, part of which was sub-granted to the Library to fund the development of a metadata-driven interface for the archive, which would both meet Library cataloging standards and provide more functionality for users than had previous Web archive interfaces.

A portion of the Election 2002 Web Archive was made available to the public in March 2003. Nearly 1,200 sites produced by House, Senate and Gubernatorial candidates were indexed, cataloged, and presented using an interface developed by WebArchivist.org. A second, research-based interface to the collection was created by WebArchivist.org on its PoliticalWeb.Info, www.politicalweb.info/home.html, site. Via this interface, visitors to the Election 2002 Web Archive can currently search campaign sites by five fields in addition to those provided by the MINERVA site. In addition, a user can access the Web archive records of pre-selected sets of URLs that demonstrate specific characteristics and thereby illustrate key findings from the analytical reports on Web campaigning generated by the WebArchivist.org research team.

5 Challenges Associated with Initiating Thematic Web Archiving

Several challenges face those interested in initiating a thematic Web collection. The response to these challenges is ideally encapsulated in the collection policy statement directing the collection activity. The challenges we found include: identifying the seed URLs for the collection activity, specifying the depth of crawling on sites represented by the seed URLs, establishing a policy for following links outside the sites represented by the seed URLs, and determining the frequency and duration of collection activities. In addition, thematic Web collections pose unique challenges for the Library of Congress

with the following: verification of collected materials; securing permission from site producers for public access; cataloging items in the collection; and providing public access to the collection.

5.1 Identification

A critical step in the collection of a thematic Web archive is the identification process for the initial seed URLs. These seed URLs are likely to be the "home pages" of sites of particular interest. A process for identifying these sites, and maintaining a continuously accurate list of these sites, should be specified as part of the collection policy statement. In addition, the collection policy statement for thematic Web archives involving unfolding events should specify a process to accommodate seed URLs that emerge during the collection period. The robustness of this process will be directly linked to the claims that can be made about the comprehensiveness and representativeness of the archive.

Once identified, the seed URLs need to be evaluated from a technical perspective to ensure that each represents the ideal starting point for a site of interest, as well as to potentially narrow the breadth of the site to be collected. Decisions about the appropriate starting page for a given site, as well as the sections of the site to be excluded from the collection are made easier with inclusion of general rules in the collection policy statement.

Finally, the list of seed URLs may need to be maintained during the collection activity, depending on the time frame specified in the collection policy statement. If the seed URLs are intended to be either a comprehensive list or representative sample of a particular genre of Web sites, and the time frame specified requires more than one crawl, it may be necessary to verify the list prior to each periodic crawl.

5.2 Internal Depth

Once the URLs are selected, or identifying the front pages of sites to be collected, it is necessary to specify the "internal depth" – that is, the number of links to follow from seed pages to pages associated with domains of the seed URLs. Internal depth can range from zero, indicating that no links are to be followed, and essentially requesting a crawl of pages rather than sites – to infinity, asking that all links that can be found be followed. Depth can be modified within sites by excluding certain URLs from collection activities, marking off or masking certain pages from the archiving process.

Once specified, the depth associated with each seed URL may need to be reconsidered in light of archiving activity and other considerations. A depth set too high may yield a collection larger than is desired. A depth set too low may yield a collection far narrower

than is desired. Finally, the depth associated with seed URLs may, by policy, be established as a constant across each seed, or may vary from site to site.

5.3 External Depth

A separate issue concerns the "external depth" – the number of links to follow from seed pages to pages not associated with domains of the seed URLs. External depth can range from zero, indicating that no links are to be follow, essentially requesting a crawl of specified sites only – to infinity, asking that all links that can be found be followed. Setting external depth to infinity would initiate a broad-based crawling strategy. A typical strategy might be to set external depth to one, requesting that the pages and page requisites linked from the sites of interest be crawled, but that links from those pages be ignored.

Managing external depth may also be a dynamic process, especially if the collection policy anticipates an external depth that is established independently for each seed URL. A depth set too high may yield a very large collection, and increase the chance that material not related to the theme or subject is collected. A depth set too low may keep relevant material from being included in the archive, and frustrate the objectives of the archivists or researchers responsible for the collection.

5.4 Periodicity

When creating an archive that contains material to be collected over time, it is necessary to specify the frequency and duration of collection. The periodicity of collection can be set across all seed URLs, or can be independent attributes of single or groups of seed URLs. Further, some pages within the sites to be collected can be archived more frequently than other pages – front pages, perhaps, collected more frequently than pages deeper into the sites. Sites or pages can be collected at any level of periodicity – from every few minutes to every few months. As it is not currently possible to accurately detect how often a given page on a Web site changes using automated crawl techniques, it is necessary to recrawl the pages or sites at regular intervals in order to allow future users to ascertain the change patterns within a page or site. Generally, periodicity is a function of expectations about changes in Web sites, the need to document the frequency and character of the changes (or lack of changes), and availability of resources to create those archives. Finally, in an event-based archive, periodicity may increase as the date of the event of interest draws closer, and taper off as the event becomes further away.

5.5 Verification

Because the value of thematic Web archives is the capture, annotation, indexing and preservation of Web sites considered useful in serving the current or future informational needs of anticipated users, the optimized goal is to capture the Web site as deeply and completely as possible.

Programmatically, Web crawler technology focuses its processing power on capturing objects. There is a need for analytical tools that can analyze Web site structure to both identify best access points for Web site captures and provide post analysis crawl data. Post analysis crawl data should allow Web analysis of site structure, content (i.e. MIME types, versioning), and linkages. Additional tools could also be useful to conduct comparisons between what was captured and what actually existed, and explain the differences observed.

Additional manual verification is also needed to ensure accuracy in the indexing processes, and to ensure that the archived pages can be properly reconstructed for display.5.6 Permissions

Prior to collecting a U.S. Web site, the Library of Congress currently contacts the web site owners regarding the Library's selection of their Web sites for inclusion in the national collections and preservation for posterity. In order to allow off-campus access to the archived web site, the web site owner must provide the Library with permission to display. The method of obtaining permissions is via email contact. Identification of the appropriate point of contact is critical in the expectation that an affirmative response will be required to allow for off-site access to any materials crawled. Getting permissions is difficult. Initial use of generic "mailto" addresses such as Webmaster@ and info@ have been abandoned for ineffectiveness. Presently, the Library finds it best to identify the appropriate email address and points of contact when a Web site is being reviewed for possible selection for inclusion in a collection, i.e., at the collection development stage. The current email trend to use a "challenge-response" system to limit spam will further complicate email delivery of notification and permission requests.

5.7 Cataloging

Processing or cataloging non-traditional materials, such as Web sites, by way of traditional means has already proven to be inefficient. Traditional cataloging tools such as Anglo-American Cataloguing Rules, 2nd edition (AACR2) and the Machine-Readable Cataloging (MARC) metadata scheme for description have undergone several changes since libraries realized the need to catalog these volatile resources. In an earlier aspect of

the Web archiving project, the Library selected, cataloged, and archived a small number of individual Web sites, one-by-one. Now, as the Library collects thematic Web archives (e.g. September 11, Election 2002), collection level AACR2/MARC catalog records for each theme is created in order to represent these items in the Integrated Library System (ILS). Cataloging of individual sites is a time-consuming effort, and automatic processes must be balanced against human cataloging to ensure a proper level of access for potentially massive amounts of data. Because each thematic archive may include thousands of sites, not all sites will be cataloged. Web archive projects are a great testbed for MODS and will assist the Library in answering and solving some of the challenges:

-Preliminary metadata may be enhanced when integrated with the Library's catalog at a later point.

-Compatibility with library standards will make it more efficient in the creation of records.

-MODS offers flexibility in terms of how specific the markup can be (e.g. the cataloger can subfield the elements of a subject heading or just use an LCSH string).

-MODS has potential use as an extension schema for METS (Metadata Encoding and Transmission Standard), which is an encoding format for descriptive, administrative, and structural metadata for textual and image-based works. This is a Digital Library Federation initiative and the maintenance agency is the Library of Congress Network Development and MARC Standards Office (NDMSO). METS attempts to package together these different forms of metadata, which are essential in a digital repository. This means that the MODS record could provide the descriptive metadata, which then gets packaged with the administrative and structural metadata in a future repository. (More information on MODS can be found at http://www.loc.gov/standards/mods/ and METS at: http://www.loc.gov/standards/mets/)

-MODS elements used may be viewed at the following URL: http://www.loc.gov/minerva/collect/elec2002/mods-elements.html.

5.8 Storage, Maintenance and Preservation

Typical sizes of thematic collections are from .5 TB to several TB in size. It is important to have a way of storing the data that can be easily scalable and accessed. Internet Archive uses standard IDE commercial hard drives, which are produced in large volume, cost effective and easily available from many sources. This has allowed Internet Archive to build collections quickly with no need for additional software or support, but by simply purchasing hard drives and standard Intel based PC's as servers.

Access to the collection is particularly important to determine if what you intended to collect was actually achieved. Broader access will provide invaluable qualitative analysis of the collection in terms of value of content and areas inadequately crawled. Maintaining a Web archive becomes a bigger challenge as the archive grows. For smaller archives, when using standard hardware, it is easy to use non proprietary monitoring tools to insure

continued access to the collection and manage equipment failures as well as the degradation of data over time on a given digital storage media, or "bit rot".

For smaller thematic collections, it has been the practice of the Archive to have multiple copies in different geographic locations. Therefore if there is a catastrophic failure, back ups of the collection exist in other places. Long-term preservation of the data for Web collections and the issues of migration vs. emulation have not yet been addressed or implemented for Web material at this stage.

5.9 Providing Access

For thematic collections it is typical for site owners to be contacted prior to collection regarding the selection of their Web site for purposes of preserving the historical record. In the past an email has been sent, and if the owner replies they do not want to provide access to their site, access is removed by removing reference to it. The "wayback machine" which is the onsite tool that provides access to Web sites captures uses the Index file to find and retrieve the archived Web page. If reference to that page is removed, the "wayback machine" cannot locate and retrieve the page. For the broader collection, it is only when a site owner contacts the Internet Archive directly will the site be removed from the Index. In few cases does it get removed from the actual Archive itself.

The Library of Congress will allow onsite access to any web archive collections that it has commissioned or will commission. However, as opposed to the "opt-out" nature of the Internet Archive access, the Library has instituted a policy that the web site owner has to "opt-in" to off-site Internet access to the web site.

6 Value of Thematic Web Archive v. Broad-Based Web Crawl

The concluding part of this paper comments on the value of thematic Web archiving, in comparison to broad-based Web crawling. These comments are presented from the perspective of a librarian, an archivist, and a scholar seeking to use collections to support scholarly research.

6.1 Library Perspective

The Library's focus on the creation of thematic Web archives will provide an archive of usable and accessible material for future researchers. Additionally, creation of these archives provides the multi disciplinary Web archiving team of Library staff representing

cataloging, legal, public services, and technology services the opportunity to develop experience and skills in evaluating, selecting, collecting, cataloging, providing access to, and preserving these materials. As with any format, the cost of the work and the requirements of serving, cataloging, storing, and preserving must be considered in the decision to collect Web content.

And, of course, it would be near to impossible, given current technology and limited web site provenance information, to collect the "U.S. web domain", and that would be broad-based collecting at the U.S. national level. The European national libraries are challenged to define and collect their respective web domain. It is estimated that U.S. domains comprise almost half of the web today. If the U.S. domain could be identified, the Library may not want to collect and store a broad-based collection of web because of the resources costs. However, future technology shifts may allow the Library develop a collection plan that would focus on a broad -based crawl to capture the U. S. web domain, or more.

6.2 Archivist Perspective

There is value in doing both broad based crawling and thematic crawling, and the two methods are not mutually exclusive. Creating thematic archives can ensure in-depth coverage of relevant sites to a specific theme or topic, which might have been missed doing a broad crawl. However, broad crawling may provide a curator with a collection of previously unknown URLs related to the topic area. Having the ability to do both broad based and thematic Web archiving results in the most complete Web archive using technology available today.

The full merits of broad based archiving may not be known for many years to come; however the potential repercussions of not archiving are high. Content published on the Web is time dependent, and therefore if not collected today may be lost forever, since typically no analog copy exists. It is difficult, if not impossible, to determine what information published will be of historical importance one hundred years from now. Even if one did have the capability, the resources required to select from over billions of pages of continually changing content would be overwhelming.

The technology is available today to do archiving on a global scale cost effectively. Bandwidth at 100 mb/sec is easily attainable at a fraction of the cost compared to several years ago. Crawlers are capable of collecting over a billion pages per month. Standard hardware using non-proprietary operating systems can be used, which can lower the cost per terabyte to under $2000.

Of course there are limitations and risks associated with doing large scale crawling. Large-scale crawlers today are only capable of doing "snapshot" crawls, which means there may be gaps in content collected over a time period. All crawlers use some type of prioritization to determine what order to crawl URLs and these may or may not be reflective of the most relevant content from an historical perspective. To date little research has been done on the best methods for emulation or migration of these large datasets and therefore the longevity of the collection is at risk as technology continues to change.

6.3 Scholarly Perspective

One critical distinction between thematic and broad-based archiving activities, from the perspective of those engaged in seeking to use social scientific techniques to complete retrospective analyses of the Web, is related to the issue of representativeness and sampling. Assuming that a "snapshot" of the entire Web at a single point in time is theoretically and practically impossible to compile, any archiving activity will necessarily yield only a selection of all possible Web objects that were available at the time of archiving. The criteria for selection or inclusion in the archive represent the method of sampling. If the sampling method is not based on robust and scholarly sampling methodologies, it is not possible for researchers using the archive to state, with any confidence, that the objects included in the archive are representative of all the objects that could have been included, and thus that findings based on the archived sample are generalizable. Similarly, thematic archiving activities are more likely to yield repeated captures of the same objects, thus facilitating robust over-time analyses. Broad-based crawls yield archives that may fail to provide the rigor required for scholarly analyses of developments on the Web around specific events or themes.

As a unit of analysis for scholarly studies of the Web, the notion of a Web sphere has been conceptualized not simply as a collection of Web sites, but as a hyperlinked set of dynamically yet systematically defined digital resources spanning multiple Web sites deemed relevant or related to a central theme or "object." (Foot and Schneider 2002; Foot, Schneider et al Forthcoming). The Web sphere that develops around a scheduled event, such as an election, may be more fully defined anticipatorily, than Web spheres that develop around unforeseen events such as those of September 11, 2001 or the Columbia space shuttle explosion. Scholars employing methods of Web sphere analysis delimit the boundaries of a Web sphere by a shared object-orientation and within a temporal framework— criteria that are consonant with those employed in thematic Web archiving. However, from a scholarly perspective, broad-based crawls provide important contextual information for analyses of a particular Web sphere, and excellent opportunities for researchers completing other types of analyses such as longer-term historical studies of a particular site or genre of sites, or surveys of broader-based Web phenomena.