

**Library of Congress
Election 2002 Web Archive Project**

FINAL PROJECT REPORT

Submitted by

Research Foundation of the State University of New York

on behalf of the

**SUNY Institute of Technology
WebArchivist.org**

**Steven M. Schneider
Principal Investigator**

April, 2004

Table of Contents

Executive Summary	3
Introduction	6
Project Background.....	6
Purpose of This Report.....	7
Project Expectations.....	9
Planning.....	9
Acquisition and Collection Development	9
Notification.....	10
Access.....	11
Quality Control	12
Organization - Metadata	12
Access.....	13
Project Management.....	14
Deliverables	14
Project Results	15
Planning.....	15
Collection Development.....	15
Collection Acquisition.....	19
Quality Control	21
Identification of Producer Data for Notification Purposes.....	22
Development of Metadata	24
Development of Web Application	27
Access to the Collection	30
Conclusions and Recommendations	33
Planning.....	34
Collection Development.....	35
Collection Acquisition.....	41
Quality Control	43
Identification for Producer Data for Notification Purposes	44
Development of Metadata	45
Development of Web Application	46
Access to the Collection	47
Listing of Appendices	48
Appendix A – SUNY~LOC Cooperative Agreement.....	49
Appendix B – Collections Policy.....	50
Appendix C - Crawl Request History (Overview of Electronic Appendix).....	51
Appendix D - Crawl History Detail by URL (Overview of Electronic Appendix)	52
Appendix E - Crawl History Summary by URL (Overview of Electronic Appendix). ..	53
Appendix F - Notification Summary by URL (Overview of Electronic Appendix)	54
Appendix G- Derivation of Metadata Fields by Producer Category (Overview of Electronic Appendix)	55

Appendix H- Description of Fields in XML Data Files by Producer Category
(Overview of Electronic Appendix)..... 56
Appendix I - Sites in Archive (Overview of Electronic Appendix) 57
Appendix J – Illustrated Screenshots of Web Application..... 58
Appendix K - Compressed Archive of Files (Overview of Electronic Appendix)..... 62
Appendix L: Web Links to Election 2002 Web Archive, as of March, 2004 63

Executive Summary

WebArchivist.org, a collaborative project at the University of Washington and the SUNY Institute of Technology, initiated the Election 2002 Web Archive project with the Library of Congress. Using funding available from a grant to the University of Washington from the Pew Charitable Trusts, the Library contracted with the SUNY Institute of Technology to develop and implement tools and processes to support the identification, acquisition, collection and cataloguing of Web sites for inclusion in the Election 2002 Web Archive, to collect and process metadata associated with the sites in the collection, to identify site owners for notification purposes and to create an interface allowing users of the archive to identify sites of interest.

This report is intended to serve as an overall summary of the project, and as a guide to future web archiving initiatives. The report summarizes the project expectations, the processes of the participants, and the project results. Most importantly, it identifies specific recommendations for future archiving efforts. Additionally, specific documentation of critical project components, including crawl histories and the structure of metadata databases, are provided in printed and electronic appendices.

In the conclusions and recommendations section, specific suggestions are provided with respect to each of the project components. These suggestions serve as the overall recommendations to the Library for future archiving projects:

Project planning:

- Build in a planning and testing phase, involving all collaborators in the project, to run for at least 60 days prior to project implementation.

- Include in this phase a test of all project components, including identification, acquisition, verification and cataloging.
- Use the planning phase to identify and resolve significant differences of perspective among project collaborators.

Collection Development:

- Implement a robust system to support collection development in which subject specialists are charged with the responsibility of providing basic metadata and object definition for each desired collected object.
- Define producer types as explicitly as possible, relying on official, government sources whenever possible, clearly defining a universe of actors within which Web sites will be sought and identified.
- In cases in which the universe of Web sites is too large to comprehensively collect, or in which the universe of potential actors is too large to support comprehensive identification of potential Web sites, develop an explicit sampling strategy in the collection policy for each site category.
- In cases in which sampling is necessary, develop a strategy that will yield the broadest collection of Web sites with the most efficient identification process. Consider employing political variables, such as campaign competitiveness, to guide election-related archives.

Collection Acquisition:

- Implement a robust system to support acquisition, providing capability for acquisition specialists to translate definitions and parameters provided by collection specialists into technical terms as specified by a collection agent for archiving, verification and correction.

Quality Control:

- Implement a robust system to support quality control, providing capability for verification specialists to determine the quality of archived objects, and providing feedback to acquisition specialists for correction of acquisition definitions and management of collection agents.
- Include in the project description a specific rate of verification required for archived objects.

Notification:

- Develop a system for collection specialists or other designated analysts to gather and enter information necessary for notification at the collection development phase, and link notification information to the acquisition records.
- Develop a system for notification email and other contacts to be recorded as part of the acquisition record for archived objects, and for this information to be provide the basis for access permissions in the object metadata.

Development of Metadata:

- In relatively small collections (fewer than 5,000 catalogued objects), do not rely on machine collection of metadata for catalog fields.

Development of Web application:

- Include detailed specifications for Web application to serve as interface to archive in overall project plan.

Provide Access to the Collection:

- Expand publicity of the archive by the Library to facilitate additional public access to the collection.

Introduction

Project Background

WebArchivist.org, a collaborative project at the University of Washington and the SUNY Institute of Technology, initiated the Election 2002 Web Archive project with the Library of Congress. Using funding available from a grant to the University of Washington from the Pew Charitable Trusts, the Library contracted with the SUNY Institute of Technology to develop and implement tools and processes to support the identification, acquisition, collection and cataloguing of Web sites for inclusion in the Election 2002 Web Archive, to collect and process metadata associated with the sites in the collection, to identify site owners for notification purposes, to create an interface allowing users of the archive to identify sites of interest, and to provide a report on this project to facilitate future Web archiving initiatives.

This project is an example of a thematic Web archive. 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.

The collection of Web materials for the 2002 Election 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, catalogued, 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 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.

Purpose of This Report

This report is intended to serve as an overall summary of the project, and as a guide to future web archiving initiatives. The report summarizes the project expectations,

the processes of the participants, and the project results. Most importantly, it identifies specific recommendations for future archiving efforts. Additionally, specific documentation of critical project components, including crawl histories and the structure of metadata databases, are provided in printed and electronic appendices.

Project Expectations

Planning

The Agreement between the Library and SUNYIT (Appendix A) anticipated that both organizations would work together to produce detailed plans and timetables for each phase of the project. The plans were expected to address the technological, policy and procedural issues of the undertaking. The planning phase was expected to afford sufficient time for testing software applications and processes. The Agreement anticipated that the Library would provide SUNYIT with the technical information and specifications necessary for SUNYIT to work collaboratively with the Library to accomplish the goals of this project.

Acquisition and Collection Development

The Library was to establish a contract with a website acquisition agent. The contract was to require that the acquisition agent liaise with SUNYIT. SUNYIT was expected to work with the acquisition agent to send and update uniform resource locators (URLs) and corresponding acquisition profiles, e.g., add/delete sites, change frequency, depth, breadth, etc. SUNYIT was to develop automated interfaces with the acquisition agent to facilitate these processes.

The Agreement anticipated that the Library would provide SUNYIT with a collection development policy statement for the Collection. The collection development policy statement was to identify both particular sites to be collected and profile desired additional types of sites (e.g., all candidate sites, party sites, interest group sites,

government sites, etc.); define acquisition parameters for each site/type, (e.g., collection frequency, depth and breadth); and determine the total number of sites to be collected, within resource parameters. In addition, SUNYIT was expected to recommend additional sites to be collected and, as appropriate, suggest changes to site profiles. To the maximum extent practicable, SUNYIT was expected to identify and include election primary sites (state and national) for use in the testing phases of this project. SUNYIT was to identify sites that met the collection development policy, and maintain the list of sites being collected, with the parameters associated with each site. On a biweekly basis, SUNYIT was to provide the Library with a list of sites being collected for the Library to review the list and update it as necessary. Reporting was to be more frequent when the election neared and as events warranted.

Notification

The Library's website acquisition agent was to provide a notice to each site being harvested for inclusion in the Collection. The notice was to explain the Library-sponsored web collection endeavor, and request that sites that did not wish to be included in the publicly-accessible Collection respond to a password-protected Library of Congress email address that was to be made accessible to SUNYIT. SUNYIT was to maintain lists of the sites' responses, identifying, for example, sites that decline to be in the Collection altogether and/or sites that opt out of allowing public web access as part of the Collection. As warranted, SUNYIT was to develop additional lists based on other types of website responses. SUNYIT was to provide these lists to the Library on a biweekly basis. The Library was to determine whether the Library or SUNYIT would contact

these sites to discuss their consent or lack thereof. The Library's website acquisition agent was to continue to collect all sites, regardless of a site's positive or negative response to the notice. The Library was expected to make a final decision on providing access to a particular site in the Collection will be a decision for the Library. It was anticipated that, as a part of the notice process, some emails would be returned to the Library's acquisition agent as undeliverable, due to incorrect addresses, web site down time, or other factors. The acquisition agent was to forward such "undeliverable notices" to a Library of Congress email address accessible to SUNYIT, for SUNYIT to follow up. For each undeliverable notice, SUNYIT was to make its best efforts to obtain correct contact information for the website, and transmit that information back to the acquisition agent, who was then to re-send the notice to the website. SUNYIT was to provide the Library with a list of undeliverable notices, and for each site on the list note whether the correct address has been located and transmitted to the acquisition agent.

Access

During the collection phase and while the acquisition agent was processing the Collection for delivery to the Library; the Library was expected to work with SUNYIT and the acquisition agent to ensure that SUNYIT had access to the archived websites through the acquisition agent. After the acquisition agent had transferred the website files and indexes to the Library, SUNYIT was to continue to have access to these materials through the term of this agreement. Access to the Collection was to be provided to SUNYIT exclusively for the purposes and uses outlined in the agreement.

Quality Control

The Library's acquisition agent was expected to perform basic quality control to confirm that the web acquisition process was working properly, e.g., that the correct URLs were crawled, that data was actually acquired from the sites, and that sites that might have "timed out" or might have been "not responding" were re-crawled. The acquisition agent was to provide quality control reports to SUNYIT and the Library that provided error information, such as error codes. SUNYIT was to develop and implement a quality control process to check that the content of the Collection matches the collection development profile and that the indexes generated by the acquisitions agent worked properly. The SUNYIT quality control review was to be used to ensure that SUNYIT's indexes and metadata were accurate. SUNYIT was to report to the Library and the acquisition agent any quality control issues that were identified.

Organization - Metadata

The Library and SUNYIT were to develop various levels of metadata, and specifications of metadata for each level, to be created by each respective party for the sites in the Collection. Level 1 metadata was to be machine-harvested generated by the acquisition agent or by SUNYIT. Level 2 metadata was to developed by SUNYIT and was to include those fields necessary to satisfy the Library's cataloguing objectives. Level 3 enhanced metadata was to be identified by the Library and to be developed by SUNYIT and applied to sites selected by the Library. Level 4 metadata was to be developed by SUNYIT, and offered to the Library at additional cost as funds were available, and as the Library wished to include in its own collection.

SUNYIT was to consider developing software tools to automate the generation of metadata. The Library anticipated that this would be possible with respect to some metadata, but that other metadata would continue to be created manually. SUNYIT was to test any proposed automation, and consider developing an interface for use by its student workers to view websites and create metadata (e.g., with pick lists for structured vocabulary or other metadata requirements).

The Library and SUNYIT were expected to cooperate to develop appropriate training materials for the workers who were to create the metadata under this Agreement. The training materials were to cover the creation of Level 2 metadata, and possibly the creation of Level 3 metadata. If the Library decided to require creation of Level 3 metadata for the Collection, the Library was to provide an instructor who would travel to a location designated by SUNYIT to conduct training on applying Level 3 metadata. SUNYIT was to identify and retain students or other capable workers to create the metadata required under the Agreement. The metadata was not to be embedded in website pages in the Collection, but rather be contained in a separate a database that linked/pointed to the websites in the Collection.

Access

SUNYIT, the Library, and the acquisition agent were to develop processes to ensure that SUNY would have access to the complete Collection (both to sites that were to be available publicly via the web and to sites that may not) while the work was in-process on a day-to-day basis and for post-processing. At a minimum, SUNY was to have “technical access” to the complete Collection during the election cycle on a next-

day basis. Access was to be available for a period not to exceed the completion date of this agreement. The Library and SUNYIT were to work out specifications for a software application for search, retrieval and display of materials in the Collection. The application was to be provided by SUNYIT and integrated into the Library web interface environment. The Library was to host the Collection on the Library's website or on another site controlled by the Library.

Project Management

SUNYIT was to provide monthly progress reports to the Library, maintain documentation on all stages of the project, and deliver in electronic and print formats the final report, with documentation, analysis and recommendations to the Library.

Deliverables

SUNYIT was to deliver a database of sites included in the Collection, with acquisition parameters for each site (e.g., associated profile, frequency, depth, breadth, response to harvesting notice, etc.), a database with site-specific metadata and related documentation, copies of software applications developed for this project, and documentation for use by the Library. SUNYIT was also to provide a detailed final report tracking the development and implementation phases of the project. The final report was to describe the project in such detail as to serve as a step-by-step guide for the Library and other institutions to use for future web collection projects.

Project Results

Planning

Nearly all of the planning phase of this project was consumed by contract negotiations between the various project partners. Final agreements among the partners were not signed until mid-June 2002 – three months later than initially anticipated. Opportunities for developing and testing an approach to site identification, acquisition, collection and verification were, unfortunately, limited to less than one month prior to the start of the archiving process. Essentially, as soon as agreements were completed among the various partners, it was necessary to begin the acquisition and collection tasks. This delay had serious negative repercussions for the entire project, and has in the long term resulted in some less-than-satisfactory outcomes, discussed in the results section below.

Collection Development

The collection was developed following the Collection Policy specified by the Library (Appendix B). The collection plan for the Election 2002 Web Archive focused on candidate, campaign, party, government, press, advocacy group and citizen sites related to the 2002 federal and gubernatorial collections. Additionally, mayoral campaigns for major cities were included in the collection plan.

Collection development was distributed by site producer type between the Library and WebArchivist. The Library assumed responsibility for identifying civic and advocacy, press, government, and public opinion sites. WebArchivist identified candidate, citizen, and party sites. Sites identified by the Library were transmitted to

WebArchivist via an email / spreadsheet system during the two months following project inception. Additional sites were added to the collection development process as identified. In addition, a set of pages (not sites) was identified during an experimental “Election Day” period, with an emphasis on specific pages thought likely to change or update during the 24 hours prior to and after Election Day. However, no regular processes or tools were created by WebArchivist for the identification and validation of URLs to be harvested into the collection. Between June 2002 and November 2002, a total of 2,763 unique domains were included within URLs identified for the collection prior to Election Day.

Advocacy groups were identified by the Library using an internal process, and provided to WebArchivist. Sites were to have been identified as financial contributors registered by the Federal Election Commission, partisan groups, and additional categories from sources such as the Washington Information Directory. Civic groups, including national and state level non-profits with election or voting emphasis, were also identified by the Library and provided to WebArchivist. A total of 663 URLs identified as civic and advocacy groups were forwarded to the Internet Archive to be crawled. Systematic Web-based searching for sites to be considered for the collection was neither anticipated nor undertaken by WebArchivist.

Site produced by candidates were identified by WebArchivist using a web-based search strategy. Following the Library’s collection policy, the races under consideration were identified: 436 House races (including the Delegate race in the District of Columbia), 37 Governor’s elections, 34 campaigns for the U.S. Senate, and the Mayor’s races in the 15 largest cities with elections. Using a variety of web-accessible databases,

including those from the local Secretary of State offices, as well as commercial and non-profit databases produced by the National Journal, the Washington Post, and politics1.com, individual candidates in each race were identified, and, for those with Web sites, URLs were sought. Through an ongoing process involving WebArchivist and the Library, a candidate Web site was eventually defined on the basis of two characteristics: (1) being clearly focused on advancing a particular candidate for a specific office, and (2) not identified as being produced by any organization other than the candidate. In this way, Web sites produced by political parties or advocacy groups, for example, even though advancing specific candidates for office, were not included as candidate Web sites. This process proved to be cumbersome and somewhat unwieldy, as new candidate sites were launched throughout the campaign period; candidates changed URLs during the campaign; new candidates emerged during the campaign; various sources identified different sites as associated with individual candidates; some candidates maintained multiple URLs; and some candidates sought multiple offices simultaneously. Though we cannot be certain that every single candidate Web site was identified, the number of sites included in the collection are within one or two percent of estimates by other organization such as CNN and other scholars. By November 2002, a total of 1,550 URLs had been identified as candidate Web sites.

Identification of sites produced by citizens with content related to the 2002 election proved to be a substantial challenge as well. The Library's collection policy statement identified such sites as those published by individual citizens, identified on a case-by-case basis. In consultation with the Library, WebArchivist developed a strategy based on identifying an emerging type of site known as a web log or "blog." The

identification process involved submitting search queries to the Google search engine with candidate name, state, and office sought and the word “blog” (i.e. “FL Governor Bush blog” or “GA Senate Cleland blog”). A total of 1,552 queries were submitted; up to 100 returned URLs were collected for each of the queries, yielding 84,587 total links distributed across 3,381 unique URLs; these urls were then filtered to include only those with the text “blog,” resulting in a list of 1,449 URLs. Based on the collection policy, which set a goal of identifying 100 citizen sites, we selected the 100 blog sites that, collectively, accounted for the most links. Our analysis suggested that the returned links were concentrated in relatively few blog sites. Ranked by number of links, the first 25 sites included 26% of the links. Including the top 100 of the identified blog sites identified those encompassing 49% of the links identified by Google.

WebArchivist identified press sites following the collection policy statement, which specified including Web sites produced by seven major national newspapers (identified by the Library), the largest newspaper (by circulation) in each state capital, and an additional 100 newspapers ranked by circulation, as well as major online news sites specified by the Library.

Government sites were identified by the Library, and included the domains of the U.S. House, the U.S. Senate, the Federal Election Commission, state and territorial central government sites, and election board sites.

An Election Day identification process added URLs with 260 additional domains (it should be noted here that many of the domains were associated with domains already identified in the collection, but show as unique domains in this analysis. There is no way

to identify “related” domains given the nature of the data available). In addition, the Election Day identification process identified additional page URLs to be collected.

An analysis of these identified URLs by producer type, summarized in the table below, indicates the range of site identification activity. Although each URL identified was initially associated with a specific producer type, the producer type may have been modified in the cataloguing process, and the initially indicated producer type was not retroactively adjusted to reflect the cataloguing process.

**URLs identified for acquisition within Election 2002 Web Archive
Project by type of site producer**

Producer Type	Domains	URLs
Advocacy	663	674
Candidate	1550	1700
Citizen	142	159
Government	89	100
Party	156	158
Press	163	203
Election Day (not categorized by producer type)	260	1956
Total	3023	4950

Collection Acquisition

In order to implement the archiving strategy, the Library established a contract with Internet Archive to serve as the website acquisition agent (Appendix B). The Internet Archive established a format for requesting sites to be archived, and provided a mechanism for updating acquisition profiles. Sites and pages were collected by the Internet Archive, based on requests submitted by WebArchivist, from June 2002 through November 30, 2002.

The acquisition process used a crawl request file formatted in XML. Each request to the Internet Archive required the creation of a line of XML code. These crawl requests can be viewed historically to form a complete record of the crawl activity within the project. The acquisition process was based on crawl “buckets.” Sites were crawled in three different groups. Those in the weekly bucket were crawled once each week. URLs in the daily bucket were crawled every 24 hours. Sites in the four-hour bucket were crawled six times each 24-hour period. Most of the crawling activity used the daily bucket. There was some use of the weekly bucket, and experimental use of the four-hour bucket. Finally, the Internet Archive provided a “once” bucket for use to allow a nearly immediate crawl of a site, so that it would not be necessary to wait until the weekly crawl came around. WebArchivist routinely requested placed new URLs in the “once” bucket when adding to the crawl request files. The crawl requests in XML format, as submitted to the Internet Archive, is detailed in Appendix C. Over the course of the project, the Internet Archive executed 15 weekly crawls, 85 daily crawls, 56 four-hour crawls, and 42 hourly crawls; this crawl history is included as Appendix D. A summary of the URLs crawled during the regular crawling processes (excluding the hourly crawls) is available as Appendix E.

Analysis of the crawl request data by producer type provides an indication of the number of sites available in the archive, for consideration to be catalogued and included in the collection, for each producer type. It should be noted that the producer type associated with the crawl requests was not verified, and may include errors of categorization rectified in the actual catalog. In addition, corrections to the crawl requests (for example, changing a domain from a .com to a .org domain) could result in

two URLs being listed in the crawl request, although only one catalog record would be produced. Crawl requests of specific sub-pages within a web site (for example, the front page of the political section of a press organization) would also be listed in the crawl request file as a unique URL. Finally, pages in the Election Day crawl were not categorized by producer type. The number of unique URLs requested in the crawl by producer type is presented below.

The acquisition process was modified during the term of the collection to allow certain domains to be excluded from collection (i.e. yahoo.com) by adding an “exclude” parameter to the crawl requests. This exclude parameter was applied to the entire crawl, not to each individual URL. It was moderately successful in limiting the crawling of some domains, but was a somewhat blunt instrument that may have resulted in missing pages in the archive. Although the actual archiving process and delivery of the archive was outside the scope of WebArchivist’s role in the project, we note that the Library’s estimate of the size of the collection is 4,000 Web sites with a total size of 1.3 terabytes of data. Without a precise definition of a “Web site,” it is not possible to directly compare the number of domains identified in the WebArchivist crawl request history to the number noted by the Library. WebArchivist did not do analysis of the size of the data collected.

Quality Control

Quality control of the archived materials collected was limited during the collection period. WebArchivist, in collaboration with the Library, had developed a set of processes that anticipated having a large number of analysts simultaneously examine

pages from the archived collection. This process was designed to facilitate collaboration between the scholarly activity and the Library activity of the WebArchivist analysts. Significant time and resources were devoted to developing these processes, including the ability to properly allocate the time analysts invested between collecting metadata related to the Library catalog and collecting data related to the scholarly research.

It quickly became apparent, however, that it would not be possible to use the archived resources to complete this task. For much of August and September, 2002, the archive was not sufficiently responsive to support multiple simultaneous users. Further, in the absence of a sustained planning and testing phase, WebArchivist was not able to develop reporting and analysis mechanisms about the sites and pages being archived. By the time this data became available, the project was fully into the implementation phase. Insufficient resources were available for redirection into verification or quality control at that time. WebArchivist was thus limited to very basic verification work within the archive during the implementation phase, and the desired level of quality control was not achieved.

Following the data collection phase, the Library determined that it would examine, on its own, all archived materials. From this process, the Library made a determination concerning quality of materials, and determined which sites to include and exclude from the archive.

Identification of Producer Data for Notification Purposes

An important component of the Election Web archive project was to notify site producers that their sites had been identified and acquired as part of the archive, and to

determine their willingness to have public access to archived versions of their site provided by the Library of Congress. The Library's collection agent sent notification emails to site producers on behalf of the Library of Congress; some of these email notifications were based on contact information developed by WebArchivist.org. Sites on the crawl list were first notified by the Library's collection agent with an email message provided by the Library. The email address used for this first notification was webmaster@<crawl URL Domain>. Responses to this emailing, including "bounces" were sent to the Library who in turn forwarded them to WebArchivist.org for categorization and compilation. WebArchivist.org staff established a database of all URL Domains, the date of initial notification, the address used for initial notification, and the response to that notification. In addition the dataset included the details of the response; acceptance or rejection of inclusion in the archive, or that the message was undeliverable, and the email address used to reply to the notification effort.

All URLs with undeliverable responses were then subjected to a series of efforts to locate a valid contact for notification. The group was divided into Candidate and Non-Candidate pools, so that the Candidate group could be processed for notification first. WebArchivist.org located valid e-mails, postal addresses, or phone numbers and merged them with the notification database. Located email contacts were transmitted to the Collection Agent and The Library in smaller groups, as they were located, for a second email notification effort, and non-email contacts were provided to the Library. As responses to these additional email notification rounds were received and forwarded by the Library to WebArchivist, WebArchivist.org recorded the responses as it had established the system for the first notification round. Again acceptances, rejections, and

undeliverable statuses were recorded in the database, in association with the re-notification dates. URLs with undeliverable responses, or with no contact as yet determined, were examined again by WebArchivist.org and Library staff to determine if a valid contact could be found. Upon this further scrutiny, if emails were identified they were included in the next email notification round, or if other contact information was found it was provided to the Library for direct notification. Through this process as many as three attempts to notify site producers via email were made, as a part of seven distinct notification groups transmitted to the collection agent and the Library.

At the end of this process a final notification report was sent to the Library with information as to whether notification was attempted, when and the email addresses or other contact information utilized. The final notification report of is included as Appendix F of this report. Of the 3,046 domains included on the notification database, valid contact information was provided for 2,420 producers.

Development of Metadata

Following the division of sites into producer types, as outlined by the Collection Policy (Appendix B), metadata for each site included in the archive was developed by post-processing and collating data collected by WebArchivist.org researchers with data provided by the Library of Congress. In addition to providing core descriptive information about sites in the archive, the metadata is used to populate the database that forms the basis of an interface to the archived sites. This interface is described in detail below.

Some data, such as characteristics of the Web site or mission of the organization, was collected from the Web sites themselves. Other data was collected from external sources such as directories or listings provided on government or press Web sites. The collected “raw” data was processed following a set of rules, specific to each of the eight producer types. Across the eight producer types, a total of 277 metadata fields were produced and encoded as attributes within XML data files. The rules for the derivation of metadata fields are detailed in Appendix G. Some producer types required distinctive processes to generate metadata:

- **Candidates:** Sites were reviewed and evaluated for the presence or absence of a key set of features. These features included candidate biography, issue position statements, campaign finance information, campaign contact information, candidate endorsements, privacy policy, and youth information. In addition, information about each candidate, including party affiliation, office sought, state and district was developed.
- **Civic and advocacy groups:** Sites were evaluated to estimate the range of issues of concern to the group. Although this is clearly an imperfect approach to assessing the range of interests associated with each group, this method provided an approximation for cataloguing purposes. The Library provided a list of issues (abortion, aged, agriculture, arts, business, civil liberties, civil rights, consumer protection, ecology, education, elections, family, foreign relations, gun control, health, homosexuality, humanitarian assistance, justice, labor unions, mass media, military policy, peace, politics and government, religious groups, science, social problems, social service, taxation, transportation and urban affairs) which were then assigned, up to three issues per group. Subgroups for religious groups, racial and ethnic groups, and foreign countries were also assigned.
- **Citizens:** Sites were examined to determine if the archived site was a “web log” or was not a web log, or blog. If the term “blog” appeared in the URL, or if the site was formatted like a diary with dated journal entries, it was considered a blog.
- **Government:** Sites were examined to determine if the agency represented was federal, state or local; if the scope of responsibility for the agency was focused on elections or extended beyond elections; and if the agency or department was focused on one particular type of office (i.e. House of Representatives, U.S. Senate or Governors).

- Party: Sites were examined to determine the scope (national or state level) of the organization producing the site, and the focus of the organization on a particular type of office (e.g. governor, house or senate).

To represent the metadata elements, WebArchivist adopted the “Metadata Object Description Schema” (MODS), an XML schema developed by the Library of Congress’ Network Development and MARC Standards Office. MODS enables the creation of original resource description records, and includes a subset of MARC fields and uses language-based tags. The standard is maintained by the Network Development and MARC Standards Office of the Library of Congress with input from users. A separate XML file containing data associated with each producer type was provided to the Library as part of the WebArchivistl.org release of the EWA 2002 Web Application (see discussion below). Each site in the archive is represented by two elements – an “index” element and a “data” element – with a common identifier. The attributes in the index element are used as metadata to populate the search and search results aspects of the access interface, discussed below. All attributes in the index element were derived directly or indirectly from attributes in the data element. Attributes in the data element form the core metadata associated with each site included in the archive, and are included in the Web Archive records displayed by the interface, as discussed below. The attributes associated with both index and data elements for each of the eight producer types are named and described in Appendix H of this report. Following a selection process conducted by the Library, metadata for a total of 2,237 Web sites was developed. These sites are identified in Appendix I.

Development of Web Application

WebArchivist.org, in collaboration with the Library, developed a Web-based application providing access to the metadata and links to Web Archive Resource Pages associated with each site in the archive. The application links the various facets of the project together and provides a public face to the Web archive. Appendix J of this report provides illustrated screen shots of the application, which consists of two main components – Search and Web Archive Records.

Within the Search component, there are three separate subcomponents – Search Options, Search Path, and Search Results. Users are initially presented with an opportunity to select one of the eight producer types with sites in the collection (Appendix J: Figure 1). After selecting a producer type, users are brought to a second search page, with a unique configuration for each producer type. The configuration for each producer type is controlled by a set of configuration files; the discussion that follows is based on the configuration as delivered to the Library in February, 2004 (see discussion below). As illustrated in Appendix J: Figure 2, the search interface includes three primary subcomponents. The Search Options component provides a list of categories within which users can select attributes of interest. For example, in the candidates configuration, users could select attributes from four different categories – office, party, geographic area and name. Within the office category, users could select from four attributes – Governor, House, Mayor or Senate. Selecting a particular attribute in a category narrows the search to those records that share that value; the remaining categories and the attributes within those categories then presented to the user (See Appendix J: Figure 3). The Search Path, presented across the top of the search interface

records and highlighted in Appendix J: Figure 2, presents the path through the current search, and allows the user to return to a previous position in the search. The number of records available at each level of the search is also presented to the user. This innovative interface is especially appropriate in an archive of Web sites in which users may be searching for a specific group of sites that share a set of characteristics. For example, users interested in viewing the Web sites of gubernatorial candidates from a particular state could isolate the set of records that satisfy these conditions. The third subcomponent of the Search Interface, the Search Results section, is a listing of sites matching the selected attributes of the searched categories, as illustrated in Appendix J: Figure 2. Ten sites are presented on each page of search results, and a navigation menu to other pages is provided. The selected characteristics displayed are configurable for each producer type. One of the characteristics, usually the name of the site, is designated as the linking characteristic, and provides a link to a Web Archive Record associated with the individual sites. The Search Results can be sorted by any of the characteristics by clicking on the name of the characteristic.

The second primary component of the Web application consists of Web Archive Records, created for each site in the archive. Appendix I includes the Web Archive Record URL for each of these records. Clicking on a site link within the Search Results returns the Web archive record for the selected site. The Web archive, as illustrated in Appendix J: Figure 4, serves as the Library's front end to the archived sites. Each of these records includes an identical set of attributes:

- Abstract: provides a brief description of the contents of the web site

- Access condition: includes a statement of any restrictions placed on viewing the web site.
- Active site: the original URL of the site as archived.
- Alternate title: provides an alternate title for the site, in the event the title provided in the original HTML was not present or acceptable.
- Collection name: identifies each site as part of the “Election 2002 Web Archive.”
- Dates captured: indicates the first and last dates for which impressions were archived for each site.
- Genre: identifies the media type represented by the archive item; in this project, all items were identified as “Web site.”
- Identifier: is the URL of the site archived, and was used for identification purposes.
- Language: indicates the primary language of the Web site, identified following the practice of ISO 639-2 Bibliographic Code.
- Producer name: identifies the organization or individual represented by the website, using the form of name most commonly found.
- Subject: includes search headings using controlled keywords as specified by the Library of Congress.
- Title: extracted from the title tag on the Web site on the date closest to the election; if it was blank, the alternate title was substituted for the title.

Each Web Archive record also includes a link to the archived site, if the access conditions established for the site permit. The link to the archived site takes the user outside of the Web application developed by WebArchivist, and enters the user into the actual archive of Web sites. The Web Archive record is designed to serve as the “front door” to the archive of Web sites; scholars and others seeking to create links to archived sites within the collection are to be encouraged to create links to the enduring Web Archive record pages; the archived sites themselves may be reindexed or moved to other servers, thus rendering deep links to the collection unstable.

The WebArchivist application, presented to the Library in final form in February 2004 as a compressed archive file in a tar format <minerva_ewa_20040205222939.tar> and included as Appendix K of this report, was the primary deliverable of this agreement. The compressed archive file includes the eight primary data files, one for each of the producer types. These files are named with the file format, <producertype_data.xml> -- for example, <advocacy_data.xml>. Modification to these files will change the data displayed in both the Search and Web archive record components.

Access to the Collection

The portion of the Election 2002 Web Archiving consisting of the sites produced by candidates for Senate, House and Governor were made available to the public via the Library in March, 2003. Nearly 1,200 sites were indexed, catalogued, and presented using an initial version of the Web application developed by WebArchivist. The archive was presented as part of a half-day conference at the Library of Congress on March 3, 2003.

There were nearly fifty attendees at the conference, including journalists, scholars, and representatives from the National Science Foundation, the Congressional Management Foundation, the Pew Internet and American Life Project, and the Institute on Politics, Democracy and the Internet, as well as staff from across the Library of Congress. Key findings from the three indepth reports mentioned above were presented, along with an interactive demonstration of both the catalog and research-based interfaces for the Election 2002 Web Archive.

News reports about the conference and the archive opening appeared in a diverse array of media such as: the Christian Science Monitor, Information Week, The Seattle Press, The Advocate in Baton Rouge, LA, the News Observer in Raleigh, NC, the Modesto Bee, CA, the AP news service, the KCSM radio station in southern California, and EurekAlert, a service of the American Association for the Advancement of Science. Subsequently, dozens of libraries and academic institutions across the U.S. and internationally have created links to the archive. A selective listing of organizations linking to the Election 2002 Web Archive is provided in Appendix L.

Public access to the collection was achieved through a collaborative hosting arrangement between the Library, WebArchivist and the Internet Archive. The connection between the various servers was seamless, and included various pages using the same graphical elements. The Library hosts the “front page” to the archive (<http://www.loc.gov/minerva/collect/elec2002/index.html>), including information pages describing the archive and the Minerva project, on the Minerva Web site (<http://www.loc.gov/minerva>). Site visitors who clicked on the “Go to Election 2002 Web Archive” were directed to the Web application hosted by WebArchivist (<http://webarchivist.org/minerva/DrillSearch>). The remainder of the Web application relies on WebArchivist.org servers, until site visitors request to view an archived site. At that point, visitors were initially directed to an “Election 2002 Web Archive Resource Page” hosted by the Internet Archive; the archived pages of the sites in the collection were also served by the Internet Archive. As of August 7, 2003, visitors requesting to view both Web Archive Resource Pages and archived pages were directed to a server hosted by the Library of Congress (wasearch.loc.gov). It is anticipated that the Library

will redirect site visitors from WebArchivist.org servers to a system hosted entirely by the Library in the near future.

Access to Web Archive Resource pages is managed by a field within the _data.xml records described above. Links to Web Archive Resource pages are suppressed if this field includes any content. In this way, the Library can continue to provide access to the Web Archive Resource records while blocking access to the actual archived sites.

Conclusions and Recommendations

In April, 2003, shortly after the public launch of the candidate sites in the Election 2002 Web Archive, WebArchivist conducted two focus groups with project participants via conference calls-- one with Library staff and the second with WebArchivist technical and research team members. A set of about a dozen open-ended questions concerning perceptions of the planning and coordination processes, accomplishments and challenges of the project to date were posed to each group. Although the discussions were audio-taped, (and the tapes of the focus group with Library staff was sent to the Library), the recording quality was too poor to be of much use. Extensive notes were taken on the discussions within each group, and these were analyzed at the time in order to incorporate lessons learned during the first year in the remaining project work and again in preparation of this report. The focus group discussions revealed that the partners held congruent perceptions of several key aspects of the project, and significantly different perspectives on other aspects. Both groups agreed that the successful launch of the candidate site section of the archive with both a Library and research interface and within four months of the election was a remarkable accomplishment that had required Herculean efforts on the part of many and coordination on multiple levels— neither group wanted to attempt to repeat the same amount of work in the same timeframe again. Insights gleaned from the focus group discussions inform the recommendations that follow in this section.

Planning

We were not successful in developing an orderly planning and testing process in this project. Initial planning discussions and contract negotiations began in January, 2002, between the preliminary and final approval notifications of grant support from the Pew Charitable Trusts. At that time WebArchivist proposed and Library staff agreed in principle to a 60-day planning phase for the project, to begin in May, 2002 before actual activity had to begin in July, 2002, assuming that contracts could be finalized during the month following final approval notification from the PCT in March, 2002. The absence of a planning and testing process was caused largely by an inability on the part of the collaborating parties to reach timely agreements on the several contracts involved. Despite the fact that negotiations began in January, 2002, WebArchivist did not receive a contract from the Library until the end of July, 2002, by which time archiving had to begin immediately. No particular clause of any one of the contracts was the source of the delay; rather the delay appeared to be caused by the need for coordination between the contracts, a lack of focus, intensity, and clear decision-making channels within the Library, and the contingency of WebArchivist's ability to begin work on the project on the implementation of a finalized contract from the Library. Even after the contract negotiations were completed, there was an additional multi-week delay on the part of the Library in delivering the finalized contract to WebArchivist. .

A thorough planning and testing phase in future projects would benefit the Library by providing a clarity and direction to the project when the inevitable crossroads arise. This would be an especially valuable exercise in which to engage for future projects that involve collaborators from different professional backgrounds – for

example, those from the scholarly community and those from the Library community. The planning phase would likely allow differences in conceptualization, interpretation and emphasis to surface among collaborating organizations, thus identifying possible difficulties prior to the project implementation phases.

The necessity for future archiving projects to engage in detailed planning may potentially be lessened somewhat by the Library's experiences to date. The critical questions that would have been answered through a planning and testing process were, in effect, answered during the actual implementation phase of this project. Lessons learned concerning the necessity of robust acquisition records and clearly articulated and defined verification techniques, discussed below – while perhaps too late to benefit this particular collection – can now be embedded in future archiving projects without the need for a specific planning and testing process.

In summary, three specific recommendations are made with respect to project planning:

- Build in a planning and testing phase, involving all collaborators in the project, to run for at least 60 days prior to project implementation.
- Include in this phase a test of all project components, including identification, acquisition, verification and cataloging.
- Use the planning phase to identify and resolve significant differences of perspective among project collaborators.

Collection Development

Collection development proceeded generally as designed in this project. For the most part, the desired sites were identified and entered into the acquisition process as planned. There were some difficulties identifying specific types of sites, owing perhaps

to a lack of clarity in the definition of the producer type. However, these difficulties did not result in a significant degradation of the collection activity. In retrospect, the most significant weakness of the collection development process as developed in this project was the lack of a robust system to support this activity. Future archiving projects should require the development of a collection system for use by subject specialists.

Subject specialists should be able to use a Web-accessible system to provide basic data about Web objects that they wish to have added to an ongoing archiving project. Basic data that should be required of subject specialists includes the base URL of the object, a definition of the object, desired collection parameters, and identifying information about the producer or creator of the object. Objects should be defined as a single URL (perhaps a graphic image found on a Web page), a page URL (meaning the Web page and its requisites), or a site URL. Objects defined as site URL should be further defined by the subject specialists, including references to page links and domains with links that should be or should not be followed, and a specification of the level of links to be followed. For example, a subject specialist might fully define a Web object, in reference to a supplied base URL, as “this page and its requisites, all pages (and requisites) linked to this page according to a current search in two search engines, all pages (and requisites) linked from this page within the same domain, all pages (and requisites) linked from subsequent pages within the same domain for three levels, and all pages (and requisites) linked from this page from any domain. The system should pre-define several “shortcut” definitions (for example, all pages for three levels within this domain) for ease of use. The desired collection parameters should be specified by subject specialists (for example – all pages weekly). Information about the site producer

provided by subject specialists would be used for cataloguing and notification purposes. Other information, including definitions and parameters, would be passed to acquisition specialists for analysis and translation into technical specifications, as required by collection agents. Object definitions and collection parameters could be established at the collection policy level, which would eliminate the necessity for including these factors within the collection development system.

This system would have eliminated the transfer of many spreadsheets and other unformatted or poorly formatted datasheets with similar information, as well as standardize the metadata provided for each desired collected object prior to acquisition. In addition, it would create a clear system of responsibility for archiving, starting with subject specialists, who would assume one of their traditional roles of defining objects to be acquired by the Library.

In addition to development of a system to facilitate collection development, future projects would benefit from a more transparent collection policy, especially those focused on an electoral Web sphere with multiple actor types contemplated for inclusion. For example, the definition of “candidate” and “civic and advocacy group” were not clear enough to define, with certainty, the entire universe of sites to be considered. For example, candidates can be defined in terms of ballot status as certified by election officials; this however does not allow for write-in candidates to be included. At the same time, the provided definitions were sufficiently clear in more than 95 percent of the cases; it may be that the cost of additional clarity outweighs the benefits.

Two related issues concern the notion of sampling sites from a defined universe, and of documenting the strategy and technique(s) employed in selecting sites from an

undefined universe. This is an especially important issue for scholars, one potential audience for future archives. In many cases, archive projects that will form the basis of scholarly research will rely on a sampling or selection strategy of some sort. The sampling strategy can be either implicit or explicit. The Library community has traditionally used an implicit sampling strategy (for example, by collecting a limited sample of newspapers from the universe of all newspapers) for its non-digital collections. This practice often relies on individual expertise and may vary across genres within a single collection. The issue of sampling, however, may require renewed attention in the digital realm, particularly in collections that are intended to be useful to scholars who require an explicit sampling strategy, consistent across genres or types of digital resources, in explaining what for them is a data corpus. In this project, explicit statements about sampling, especially for sites produced by citizens and civic and advocacy groups, would be helpful additions to the collection policy statement, and would increase the value and usability of the collection for scholars.

Similarly, sampling within a Web sphere may be desirable for certain archiving projects. This may especially be the case with relatively large collections, such as those focused on national elections. While it is clear that the Library has as part of its mission comprehensive collections, there may be opportunities to collect nearly comprehensively by focusing on a narrow aspect of a Web sphere. In the case of the 2002 election, it was highly plausible that a relatively well-developed set of Web sites, including those produced by civic and advocacy groups and citizens, was more likely to emerge during the election season in reference to highly competitive campaigns than in reference to non-competitive campaigns. Previous scholarly research and analysis made clear that in

campaigns with little or no competition between candidates, media activity and public discourse (including Web sites) was considerably less than in highly competitive campaigns. One approach to sampling, suggested by the research team partners on this project, would have been to identify a fixed number of the most competitive Gubernatorial, House, Senate and Mayoral campaigns, and to have initiated extensive searching for Web sites referencing those campaigns. In a context in which the universe of electoral races was too large to permit universal searching, such a sampling strategy would likely have yielded valuable results, albeit it on a smaller sample of campaigns. Library staff determined that this approach should not be followed, due to concerns that the Library's collection policies not be guided by political considerations. However, this had the effect of restricting the range of sites included in the collection, and had an especially significant impact on the nature of citizen sites included in the collection.

A closely related issue concerns dynamic identification. The initial collection policy, as developed in collaboration with the Library, did not fully account for the dynamism in the electoral Web sphere. The process for identification of sites to be included in the collection was based on an assumption that, once identified, the URL would remain stable. Our experience taught us that this was not the case. This was especially the case for candidate sites – those identified in the summer of 2002 were, in some cases, not the URL actually used by the candidate during the campaign. In some cases, a URL from the 2000 campaign was still active, only to be replaced by a new URL later in the summer or early fall. As part of the collection policy, the dynamic nature of sites and URLs should be accounted for.

Along another line, there was also some confusion, especially with Press sites, and to a certain extent, government sites, about the nature of the item to be collected. This confusion stemmed from a lack of clarity about the definition of a “Web site.” Given that a Web site is notoriously difficult to define in precise terms, it is not surprising that this issue would have caused some difficulty in this project. Our initial operating definition of a “site” was to include the front page and all URLs within the same domain that were linked from that page and from subsequent pages within the same domain. This make identification or collection development relatively straightforward, but caused problems in collection and acquisition, as discussed below. When we attempted to rectify those difficulties by identifying specific pages within the site to use as “seed” pages, this raised issues about the nature of the collection and the collection policy itself. A very clear definition of the collection objective at the outset of the project, conceptually and operationally defining the target objective, would have clarified this issue.

In summary, four specific recommendations are made with respect to collection development:

- Implement a robust system to support collection development in which subject specialists are charged with the responsibility of providing basic metadata and object definition for each desired collected object.
- Define producer types as explicitly as possible, relying on official, government sources whenever possible, clearly defining a universe of actors within which Web sites will be sought and identified.
- In cases in which the universe of Web sites is too large to comprehensively collect, or in which the universe of potential actors is too large to support comprehensive identification of potential Web sites, develop an explicit sampling strategy in the collection policy for each site category.

- In cases in which sampling is necessary, develop a strategy that will yield the broadest collection of Web sites with the most efficient identification process. Consider employing political variables, such as campaign competitiveness, to guide election-related archives.

Collection Acquisition

On the surface, the acquisition process worked fairly well. Web sites were identified for collection, forwarded to the acquisition agent, and archived as designated times. The verification process, however, did not function effectively, and we were unable to reliably develop the necessary feedback loops from the verification to collection processes. This is the one area in which significant advances and development for future archiving projects could be most valuable. A full-scale database system, coordinated with the system for collection development discussed above, should be created to serve the acquisition and verification processes. This system should include clearly identifiable components for acquisition and verification specialists, and should serve to provide an integrated interface to collection agents.

The acquisition system should provide the capability to translate the object definitions specified by the collection policy or created by the collection specialists into technical terms as specified by the collection agent. For example, in this project, the collection agent used the concept of a “mask” URL to identify the domain hosts which were to be spanned during the crawling processes. The collection agent will specify the technical details necessary for efficient crawling of desired objects; acquisition specialists should be responsible for generating these details according to the specification. The acquisition system should provide the capability to generate, test and correct these details,

as well as provide feedback to collection specialists in situations where there is a discrepancy between the technical and operational definitions of the object.

The acquisition system should include the capability to generate full-scale acquisition records. These records, including the metadata provided by the collection specialists, would provide the basis for preliminary cataloging, archiving activity, and post-archiving verification. The preliminary catalog generated by acquisition records could serve both internal Library needs, as well as allow for the rapid deployment of archived materials into the public collection. The acquisition records would include all object-related requests distributed to the collection agent. In this way, the system would be able to generate reports about the crawl status of specific objects.

The acquisition system discussed above should provide direction to a robust verification process. This system would allow verification specialists to be directed to specific portions of the collection during the implementation phase to verify that the objects have been archived as specified. Verification specialists would be able to follow specific protocols for examining archived objects and provide feedback to the acquisition specialists about the quality of the archived objects. The acquisition specialists would use this information to both fine-tune the technical definitions of the object as well as to work with the collection agent to improve the crawling processes.

In summary, one specific recommendation is made with respect to collection acquisition:

- Implement a robust system to support acquisition, providing capability for acquisition specialists to translate definitions and parameters provided by collection specialists into technical terms as specified by a collection agent for archiving, verification and correction.

Quality Control

The quality control or verification process in the current project was almost exclusively a post-collection verification; little verification was possible during the collection phase. The collection agent was not able to provide robust access to the collection, including the capability for multiple simultaneous users, until fairly late in the implementation phase. WebArchivist had intended to use a Web-based verification process involving multiple analysts; the intention was to gather the metadata required for the catalog and verify the collection at the same time. Due to system limitations, we were required to gather the metadata from the live Web, and thus were unable to do verification during the collection process.

Quality control is a critical component of an archiving project, and must be integrated more systematically into the implementation phase. The project description should specify a verification rate at the object level. For example, the project description could require that 10 percent of all archived page objects be verified by a visual inspection process measuring the percent of page requisites obtained by the archiving process, and that this verification take place within 72 hours of the archiving activity. This type of verification requirement would allow project managers to clearly identify the resources necessary to devote to the verification process, and to ensure that funds are available to sustain this activity.

The need for this system may have become clear during the desired testing phase for the project. However, the full dimensions of the system desired have primarily come from the experience of developing, collecting, verifying, cataloging and providing access

to a large-scale collection. In summary, two specific recommendations are made with respect to quality control:

- Implement a robust system to support quality control, providing capability for verification specialists to determine the quality of archived objects, and providing feedback to acquisition specialists for correction of acquisition definitions and management of collection agents.
- Include in the project description a specific rate of verification required for archived objects.

Identification for Producer Data for Notification Purposes

Identification of producer data for notification purposes was a task fraught with difficulties and misconceptions. This process, which grew in importance to a critical level during the project, was not thoroughly considered in the preliminary project planning stages. The absence of any opportunity to test these systems resulted in the development of an ad-hoc system with undesirable consequences. In short, the process that emerged was not sufficiently robust or tied to other processes to produce the desired result.

Future archiving projects should specify the importance of developing producer data for notification purposes, and ensure that this process is viewed by all project participants as a necessary and critical component of the overall project. As discussed above, a robust system integrating collection development and collection acquisition would provide the necessary data and systemic approach to support producer notification. It is suggested that this system also include a capability to process electronic mail records sent to producers, and to incorporate responses (or lack of responses) into the records associated with archived objects. In this way, producer responses to notification can be systematically incorporated into records governing access to the collection.

In summary, two specific recommendations are made with respect to the notification process:

- Develop a system for collection specialists or other designated analysts to gather and enter information necessary for notification at the collection development phase, and link notification information to the acquisition records.
- Develop a system for notification email and other contacts to be recorded as part of the acquisition record for archived objects, and for this information to be provide the basis for access permissions in the object metadata.

Development of Metadata

The metadata developed for the collection made possible the creation of an innovative Web application providing access to the archive. This metadata, as described above, is primarily focused on the characteristics of the site producers; limited attention is given to the characteristics of the Web sites. The one exception concerns candidate sites, for which metadata about the characteristics of the sites themselves was collected.

One area of concern with respect to the collection of metadata is the notion of doing machine collection. The project anticipated collecting machine data for the title of the site, as well as the dates of collection. Analysis of the title data collected by machine indicates that such an approach is not as straightforward as first anticipated. A constructed title, based on metadata about the producer and developed by collection specialists, would have been more efficient and useful in the catalog process.

There are two specific problems associated with the machine collection of metadata. First, given the lack of standards applied by Web developers to even a basic HTML field such as title, useful data is often not accessible to machines. Second, and more importantly, the issue of non-standard characters and encoding, especially when translated to XML, causes significant problems with machine data. For the relatively few

sites included in this archive, it would have been more efficient to use constructed title based on metadata developed by collection specialists.

In summary, one specific recommendation is made concerning the development of metadata:

- In relatively small collections (fewer than 5,000 catalogued objects), do not rely on machine collection of metadata for catalog fields.

Development of Web Application

The Web application developed for the Library to provide access to the collection and the metadata about the objects included in the collection is one of the most advanced and fully-developed interfaces to a Web archive that has been made publicly available to date. The interface offers an opportunity for the Library to provide search capabilities at the site level to a substantial collection of archived sites. This interface can serve as a model for future archiving projects should it be determined to meet the Library's needs.

The interface was developed largely as a stand-alone and single-project application. There was no expectation that the Library would receive a product that could be used, off the shelf, with other Web archives. Accordingly, few resources were devoted to developing the type of application that could be easily retooled for use with other archive collections. While this is certainly possible with the software as currently delivered, extension of the Web application to other environments requires technical skills and expertise.

Absent any specifications from the Library concerning the nature of the server on which the application was to reside, the interface was designed to run in an environment with low processing demands. This restricted the development of server-intensive

searching capabilities. Further, the Library indicated that full-fledged searching capabilities would be better handled by other projects in development.

Future archiving projects could benefit from additional specifications for the Web application. In summary, a single recommendation is made with respect to the Web application:

- Include detailed specifications for Web application to serve as interface to archive in overall project plan.

Access to the Collection

Public access to the collection was accomplished according to the initial schedule for the project, in accordance with the requirement by the Pew Charitable Trusts. Access to additional sites in the collection is, of course, at the discretion of the Library.

The Election 2002 Web Archive represents a tremendous resource for those interested in Web archiving in general and for those interested in the use of the Web in the electoral domain in particular. Providing access to the portions of the archive other than candidates would allow the Library to continue to serve these two audiences. One difficulty the project had with providing access to the collection concerns publicity.

Though the fact of the archive is slowly disseminating throughout the various communities of interest, the relatively few links to the archive indicate that additional publicity could promote further access to the collection. In summary, a single recommendation is made to facilitate access to the collection:

- Expand publicity of the archive by the Library to facilitate additional public access to the collection.

Listing of Appendices

Appendix A – SUNY~LOC Cooperative Agreement

Appendix B – Collections Policy

Appendix C - Crawl Request History (Overview of Electronic Appendix)

Appendix D - Crawl History Detail by URL (Overview of Electronic Appendix)

Appendix E - Crawl History Summary by URL (Overview of Electronic Appendix)

Appendix F - Notification Summary by URL (Overview of Electronic Appendix)

Appendix G - Derivation of Metadata Fields by Producer Category (Overview of Electronic Appendix)

Appendix H- Description of Fields in XML Data Files by Producer Category (Overview of Electronic Appendix)

Appendix I – Listing of Sites in Election 2002 Web Archive

Appendix K - Compressed Archive of Web Application (Overview of Electronic Appendix)

Appendix L: Web Links to Election 2002 Web Archive, as of March, 2004

Appendix A – SUNY~LOC Cooperative Agreement

Appendix B – Collections Policy

Appendix C - Crawl Request History

(Overview of Electronic Appendix)

File Name:

Crawl Request History (.XML format)

Summary:

This file provides a history of crawl requests for each of the 4,974 URLs that were crawled during the project. The file is an XML file with the following format:

```
<seed-data>

  <crawl-info url=[url]>

    <crawl category=[producer category of url, or LIVE for Election Day
      crawl] period=[crawl bucket] stopped=[timestamp when crawl stop
      was requested] started=[timestamp when crawl start was
      requested]>
    <include url=[url]/>

  </crawl>

</crawl-info>

</seed data>
```

Appendix D - Crawl History Detail by URL**(Overview of Electronic Appendix)****File Name:**

Crawl Detail by URL (.txt format)

Summary:

This file provides crawl details for each of the 149 crawls executed during the 2002 Election Web Archive project. The file contains 92,452 lines. Each line is delimited into three elements by a pipe or vertical bar (|) representing the following:

Element	Description
1	Crawl Type
2	Crawl Number
3	URL Crawled

Appendix E - Crawl History Summary by URL

(Overview of Electronic Appendix)

File Name:

Crawl Summary by URL (.xls format)

Summary

This file summarizes crawl activity detail at the URL level. For each of the URLs crawled within the 2002 Election Web Archive project, the following detail is provided:

Column	Heading	Description
A	URL	Crawled URL
B	24h_times	Number of times crawled in 24 hour bucket
C	24h_first	First crawl in 24 hour bucket
D	24h_last	Last crawl in 24 hour bucket
E	24h_not_times	Number of times not crawled in 24 hour bucket
F	24h_not_first	First not crawl in 24 hour bucket
G	24h_not_last	Last not crawl in 24 hour bucket
H	1w_times	Number of times crawled in 1 week bucket
I	1w_first	First crawl in 1 week bucket
J	1w_last	Last crawl in 1 week bucket
K	1w_not_times	Number of times not crawled in 1 week bucket
L	1w_not_first	First not crawl in 1 week bucket
M	1w_not_last	Last not crawl in 1 week bucket

Appendix F - Notification Summary by URL

(Overview of Electronic Appendix)

File Name:

complete_notification_data_20030421 (.xls format)

Summary:

This file provides a summary of the notification attempts and results for producers of sites crawled for the 2002 Election Web Archive project. The file contains 3025 rows, each with 16 columns

Column	Title	Description
A	URL	URL Crawled
B	Domain	Root Domain
C	Notified	Contact identified and supplied for notification
D	notifiedIAPre20021004	webmaster@<domain> sent by IA
E	notifiedIA1	Email contact supplied to IA 10/25/02
F	notifiedIA2	Email contact supplied to IA 11/02/02
G	notifiedEWA3	Email contact supplied to Library 10/25/02
H	notifiedEWA4	Email contact supplied to Library 11/02/02
I	notifiedEWA5Cand	Candidate contact supplied to Library 12/10/02
J	notifiedEWA5NonCandEmail	New candidate email contact supplied to Library 12/10/02
K	notifiedEWA5NonCandContact	New non-candidate contact supplied to Library 12/10/02
L	notifyMayoral	Initial mayoral candidate email contact supplied to Library
M	Category	Producer type
N	Identifier	Archive ID #
O	CrawlTypes	Crawl type per Appendix E
P	FirstCrawl	Date of first crawl

***Appendix G- Derivation of Metadata Fields by Producer Category
(Overview of Electronic Appendix)***

File Name:

Derivation of Metadata Fields by Producer Category.pdf (pdf format)

Summary:

This file provides a detailed description, for each category, of the derivation of the metadata fields included in the _data.xml files for the Election 2002 Web Archive.

Appendix H- Description of Fields in XML Data Files by Producer

Category

(Overview of Electronic Appendix)

File Name:

Description of Fields in XML Data Files by Producer Category.pdf (pdf format)

Summary:

This file provides a description of fields in the XML data files for the Election 2002 Web Archive.

Appendix I - Sites in Archive**(Overview of Electronic Appendix)****File Name:**

List of sites included in Election 2002 Web Archive.pdf (pdf format)

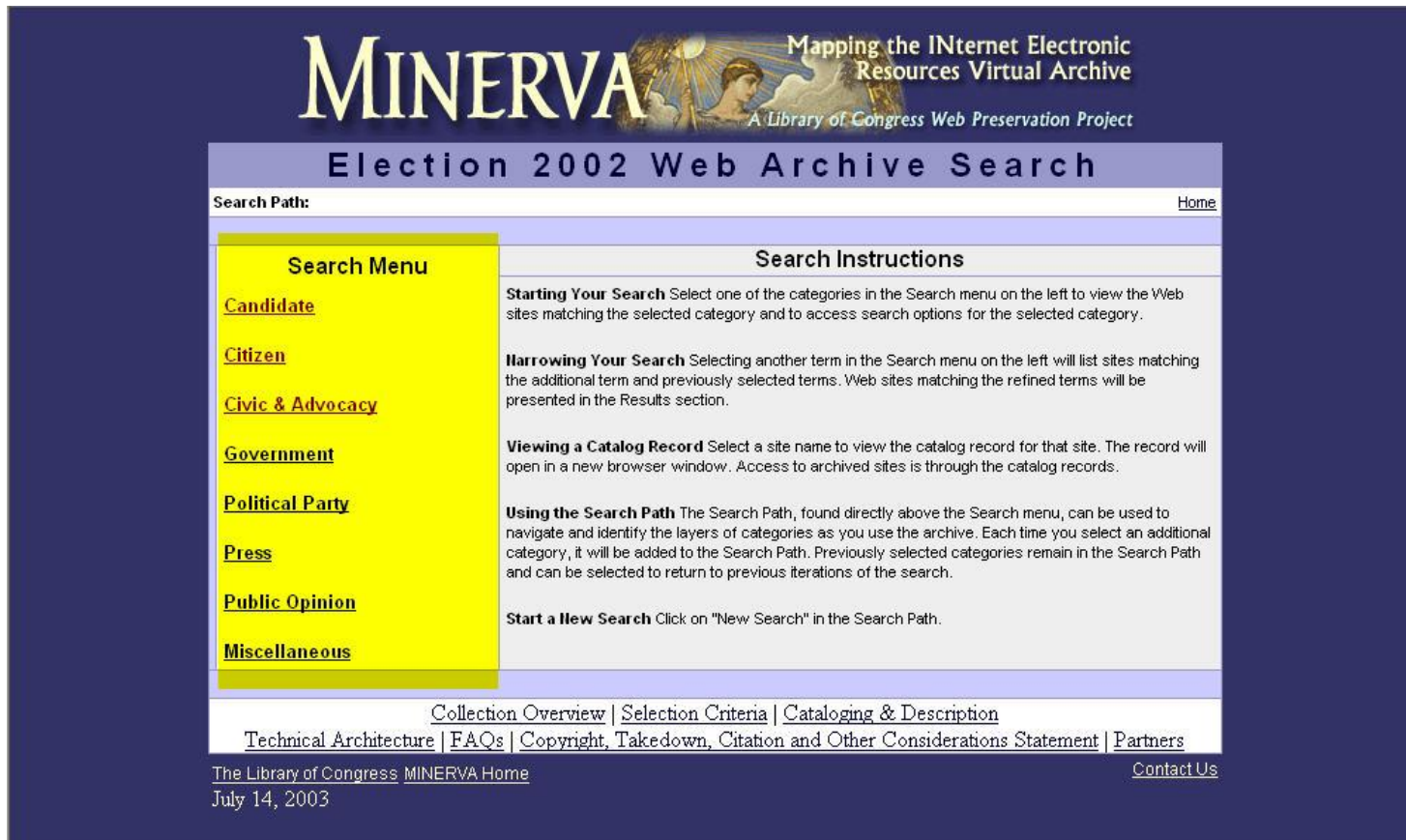
Summary:

This file provides a list of sites included in the Election 2002 Web Archive. For each site, the following data is provided:

Column	Heading	Description
1	Producer Name	Name of producer, as included in data file
2	Category	Producer type
3	Site ID #	Unique identifier; use to find Web Archive Record
4	First Capture Date	Date of first impression of site in archive
5	Last Capture Date	Date of last impression of site in archive
6	URL	Original URL of site as captured

Appendix J – Illustrated Screenshots of Web Application

Figure 1: Initial Search Page in Election 2002 Web Application



Yellow highlighted section illustrates opportunity to select from among eight producer types

Figure 2: Sample Search Interface Page in Election 2002 Web Application

MINERVA Mapping the INternet Electronic Resources Virtual Archive
A Library of Congress Web Preservation Project

Election 2002 Web Archive Search

Search Path: [New Search](#) | **Category** [Candidate] (1168) | [Home](#)

Search Options

Office
[Governor](#) [House](#) [Mayor](#) [Senate](#)

Party
[Democratic](#) [Green](#) [Libertarian](#)
[Other Third Party](#) [Reform](#)
[Republican](#)

Geographic Area
[AK](#) [AL](#) [AR](#) [AS](#) [AZ](#) [CA](#) [CO](#) [CT](#)
[DC](#) [DE](#) [FL](#) [GA](#) [GU](#) [HI](#) [IA](#) [ID](#) [IL](#)
[IN](#) [KS](#) [KY](#) [LA](#) [MA](#) [MD](#) [ME](#) [MI](#)
[MN](#) [MO](#) [MS](#) [MT](#) [NC](#) [ND](#) [NE](#) [NH](#)
[NJ](#) [NM](#) [NV](#) [NY](#) [OH](#) [OK](#) [OR](#) [PA](#)
[RI](#) [SC](#) [SD](#) [TN](#) [TX](#) [UT](#) [VA](#) [VI](#) [VT](#)
[WA](#) [WI](#) [WV](#) [WY](#)

Name
[A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#)
[O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

[All Candidates](#)

Results

Name	Geographic Area	Party	Office
Abbott, Jim	SD	Democratic	Governor
Abbott, Phillip	TX-18	Republican	House
Abercrombie, Neil	HI-1	Democratic	House
Abu-Ghazaleh, Maad	CA-12	Libertarian	House
Adam, Iris	CA	Other Third Party	Governor
Aderholt, Robert	AL-4	Republican	House
Adkins, Bob	FL	Other Third Party	Governor
Ahumada, Pat	TX-27	Republican	House
Akin, Ren	WY-At Large	Democratic	House
Akin, Todd	MO-2	Republican	House

Number of records: 1168
[1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [Next](#) [Last](#)

[Collection Overview](#) | [Selection Criteria](#) | [Cataloging & Description](#)
[Technical Architecture](#) | [FAQs](#) | [Copyright, Takedown, Citation and Other Considerations Statement](#) | [Partners](#)

The Library of Congress MINERVA Home [Contact Us](#)
 July 16, 2003

Yellow highlighted section illustrates drill search path; pink highlighted section illustrate drill search fields; blue highlighted section illustrates search results.

Figure 3: Sample Level 2 Search Interface Page in Election 2002 Web Application

MINERVA Mapping the Internet Electronic Resources Virtual Archive
A Library of Congress Web Preservation Project

Election 2002 Web Archive Search

Search Path: [New Search](#) > [Category \[Candidate\]](#) > **Office - House (851)** [Home](#)

Search Options		Results			
		Name	Geographic Area	Party	Office
Party Democratic Green Libertarian Other Third Party Reform Republican		Abbott, Phillip	TX-18	Republican	House
Geographic Area AK AL AR AS AZ CA CO CT DE FL GA GU HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC ND NE NH NJ NM NV NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV WY		Abercrombie, Neil	HI-1	Democratic	House
Name A B C D E F G H I J K L M N O P Q R S T U V W X Y Z All Candidates		Abu-Ghazaleh, Maad	CA-12	Libertarian	House
		Aderholt, Robert	AL-4	Republican	House
		Ahumada, Pat	TX-27	Republican	House
		Akin, Ron	WY-At Large	Democratic	House
		Akin, Todd	MO-2	Republican	House
		Alexander, Klint	KY-1	Democratic	House
		Alexander, Rodney	LA-5	Democratic	House
		Alfonsi, Phil	WI-2	Republican	House
		Number of records: 851 1 2 3 4 5 6 7 8 9 10 Next Last			

[Collection Overview](#) | [Selection Criteria](#) | [Cataloging & Description](#)
[Technical Architecture](#) | [FAQs](#) | [Copyright, Takedown, Citation and Other Considerations Statement](#) | [Partners](#)
 The Library of Congress [MINERVA Home](#) [Contact Us](#)
 July 16, 2003

Yellow highlighted section illustrates drill search path at level 2; pink highlighted section illustrate drill search fields remaining after Office has been selected into the drill search path

Figure 4: Sample Web Archive Record in Election 2002 Web Application

The screenshot shows a web archive record page with a dark blue background. At the top left is the 'MINERVA' logo in large white letters, with a small illustration of a woman's head and a sunburst. To the right of the logo is the text 'Mapping the Internet Electronic Resources Virtual Archive' and 'A Library of Congress Web Preservation Project'. Below this is a yellow rectangular box containing the record details. At the bottom of the page are three links: 'The Library of Congress', 'Close window to return to search page', and 'Contact Us', along with the date 'March 4, 2003'.

MINERVA Mapping the Internet Electronic Resources Virtual Archive
A Library of Congress Web Preservation Project

Election 2002 Web Archive Record

Title: Sherwood Boehlert U.S. Congress - Home

Alternative Title: Sherwood "Sherry" Boehlert, Republican Party and Independence Party candidate for House, New York, 24th District, 2002.

Name: **Boehlert, Sherwood "Sherry"**

Abstract: Web site promoting the candidacy of Sherwood "Sherry" Boehlert, Republican Party and Independence Party candidate for House, New York, 24th District, 2002. Includes candidate biography, issue position statements, endorsements and campaign news press releases. Site features enable visitors to sign up for campaign email list, volunteer and make campaign contributions.

Date Captured: August 9, 2002 - November 30, 2002
[Archived Site](#)

Subjects: Elections--New York (State)
United States.. House of Representatives--Elections
Republican Party (NY)
Third Parties

Language: English

Genre: Web Site

Access Condition: None

Active Site: <http://www.boehlert.com/>

Collection Title: [Election 2002 Web Archive](#)

[The Library of Congress](#) [Close window to return to search page](#) [Contact Us](#)
March 4, 2003

Appendix K - Compressed Archive of Files (Overview of Electronic Appendix)

File Names:

minerva_ewa_20040205222939.tgz (tar gzip format)

minerva-docs.zip (zip format)

minerva-docs.xml (xml format)

Summary:

These files provide the delivery and documentation of the Election 2002 Web Archive, including the data files. The file <Minerva-doc.zip> provides the documentation in HTML format, and should be consulted first. The gzip file can be opened using standard zip extractors, and includes a single tar archive, which can be extracted using standard zip extractors, including WinZip and other similar products.

Appendix L: Web Links to Election 2002 Web Archive, as of March, 2004

Source: <http://www.google.com/>.

Query: [link:http://www.loc.gov/minerva/collect/elec2002/](http://www.loc.gov/minerva/collect/elec2002/)

1. [The Telson Spur: Field Nodes -- Values \(1\): Rights & ...](#)

[Home] [Fore] [Down] [Aft] Contents, Jump, Search, Gopher, Meta, Nodes, Home, Index. Values. The Way of the Spirit. DESCRIPTION: In the search ...
www.snark.ca/pol.htm - 90k

2. [ELECTIONS](#)

ELECTIONS and CAMPAIGNS. MAJOR GOVERNMENT ELECTRONIC TITLES.
Major Web

Site: Federal Election Commission Enforces federal election laws. ...
www.swem.wm.edu/GOVDOC/election.html

3. [External web links - Leeds University Library](#)

External web links for subject: Politics and International Studies.
Broader subjects, Business, Law, Education and Social Sciences. ...
www.leeds.ac.uk/ROADS/subject-listing/service/32.html

4. [New Web Sites and Publications](#)

New Web Sites and Publications, Following is a list of sites and publications
on the World Wide Web which are newly-available within ...
home.wlu.edu/~grefed/new_web.html

5. [Campaigns & Elections](#)

Herbert H. Lehman Social Sciences Library, 300 International Affairs 420 W. 118th
St. / New York, NY 10027 (212) 854-4170 / lehman@libraries.cul.columbia.edu. ...
www.columbia.edu/cu/lweb/indiv/usgd/campaign.print.html

6. [Context](#)

Links, A presidential election campaign follows a set of familiar steps,
from the early maneuvering and testing-the-waters activities ...
www.gwu.edu/~action/2004/chrncont.html

7. [\[PDF\] DigiCULT.Info Issue 6 - ISSN 1609-3941](#)

Info A Newsletter on Digital Culture Issue 6 ISSN 1609-3941
December 2003 DigiCULT.Info 1 This is a rich issue of DigiCULT ...
www.digicult.info/downloads/dc_info_issue6_december_20031.pdf

8. [Meriam Library -- US Election Information](#)

meriam library -- government resources US Election Information. skip to contents

ReSEARCH Station Library Catalog Periodicals/Journals Subject Resources ...
www.csuchico.edu/library/gov/election_us.html

9. [Information Services & Resources: Elections & Campaigns](#)

Information Services and Resources, Library Catalog ISR Site. ...

www.isr.bucknell.edu/collection_guides/government_information/elecampa.asp

10. [Election](#)

Choose Destination. ...

www.washlaw.edu/doclaw/election.html - 4k - Mar 4, 2004

11. [PDF] [University of California/Stanford](#)

University of California/Stanford Government Information Librarians
Group Meeting March 20, 2003 9 am - 4 pm Agenda and Minutes ...

www.library.ucsb.edu/gils/GILSMin03-20-03.pdf

12. [BGSU Documents, Government Documents in the News](#)

Documents in the News. fdlp image DOCUMENTS HOME,
fdlp image LIBRARY HOME, fdlp image BGSU HOME. ...

www.bgsu.edu/colleges/library/services/govdocs/news.html

13. [Election Information by M. Finley](#)

Election Information, ...

library.csun.edu/mfinley/election.html

14. [SMSU Libraries](#)

Election 2000. ...

library.smsu.edu/resources/election.shtml

15. [Hot Topics - Election Reform](#)

Election Reform. Recommended Library Resources: Academic Search Premier
-- simply search "election reform" . The citations that come ...

www.murraystate.edu/msml/hottopicselection.htm

16. [Elections](#)

American Presidential Elections from Britannica includes election results,
biographies, Presidential documents, 1789- . Center for ...

www-libraries.colorado.edu/ps/gov/us/election.htm

17. [Politics and Elections - Web Sites by Subject - FCPL](#)

Politics and Elections. you are here: homepage > libraries & museums > library
homepage > reference resources > web sites by subject > politics and elections. ...

www.co.fairfax.va.us/library/internet/politics.htm

18. [SULAIR: Research Quick Start Guides: Elections \(US\)](#)

skip to page content | skip to main navigation, ...

www-sul.stanford.edu/research_help/res_quick_start/elections.html

19. [NOCALL - Politics](#)

Northern California Association of Law Libraries. | HOME | California Resources | Internet Resources | Publishers | Libraries | Survival |. ...
www.nocall.org/politics.htm

20. [ResearchBuzz: Government-Elections Archives](#)

ResearchBuzz! ResearchBuzz Logo Search Engine News and More Since 1998.
Get ResearchBuzz by E-Mail! Weekly Newsletter Privacy Policy. ...
www.researchbuzz.org/archives/cat_governmentelections.shtml