The UCLA Online Campaign Literature Archive: a Case Study

Gabriella Gray and Scott Martin
UCLA Library, Young Research Library
Box 951575
Los Angeles, CA 90095
gsgray@library.ucla.edu, smartin@library.ucla.edu

ABSTRACT
The UCLA Online Campaign Literature Archive has been actively archiving websites related to Los Angeles and California elections since 1998. Conceived as an extension to the UCLA Library's existing printed Campaign Literature Collection, the Archive currently contains 1138 fully cataloged and searchable websites. This paper describes the processes used by the Archive's staff to select, capture, assemble, describe, and provide access to websites.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *collection, dissemination, standards, systems issues*

General Terms
Management, Documentation, Design, Reliability, Experimentation, Human Factors, Standardization, Legal Aspects

Keywords
Political Campaigns, Web Archiving, Web Capture, Elections

1. INTRODUCTION
The Collections, Research, and Instructional Services (CRIS) department at the UCLA Young Research Library maintains a Campaign Literature Collection containing a century of printed ephemeral election materials distributed by campaigns for local, state, and federal offices and for ballot measures affecting the Los Angeles area. In 1998 a project was initiated to scan selected items from the collection and to capture and preserve copies of campaign websites for inclusion in what would become the UCLA Online Campaign Literature Archive (the Archive) at http://digital.library.ucla.edu/campaign/.

The Archive falls directly into Masanès's topic-centric project cluster "undertaken by libraries without informant networks but with dedicated staff that provide a manual verification of archived sites". [7] The Online Campaign Literature Archive is maintained by two CRIS staff members as just one aspect of their overall job responsibilities.

This paper presents a case study of a small, working web archive. It is intended for both a technical and non-technical audience.

2. SELECTION
In the weeks leading up to an election, Archive staff compile lists of candidates and ballot propositions using official state and local government websites. A variety of sources (certified lists of candidates, third party political websites, Google, newspaper articles) are consulted to find campaign websites that fall within the Archive's scope:

- Sites must be devoted to a specific election, not an ongoing organization or cause.
- Sites must be about candidates or measures which appear on the ballot.
- Elections for executive and legislative offices are collected, not judicial.
- Legislative candidates must be from districts which are substantially within Los Angeles County.
- State elections gathered include U.S. Congress, California state-wide offices and measures, and California state legislators.
- Municipal elections gathered include those associated with Los Angeles County; the cities of Los Angeles, Santa Monica, Beverly Hills, and Culver City; and any special districts (school districts, water districts, etc.) that substantially overlap those cities.

An analysis of the statistics from the 2006 California general election reveal that, of the 164 candidates falling within the Archive's scope, 105 had at least one website. An increasing number of candidates have multiple sites including blogs and MySpace pages.

As the list of sites is compiled, the individual sites are analyzed in terms of size, structure, and technical details which may impact the capture process. One of the primary findings of the Archive is that no single capture process works for all websites. The Archive uses three different web capture programs: WebCopier 4.4, HT Track Website Copier 3.33, and Offline Explorer Pro 4.6. Each has its unique strengths and weaknesses, and each also offers numerous settings which can be used to customize the capture process to fit the specific characteristics of individual sites. Web capture software programs must also be constantly upgraded to handle new web technologies and trends. Before each election

IWAW'07, June 23, 2007, Vancouver, B.C.

Archive staff must therefore download and review the latest software upgrades for each program. Section 4 presents a summary of the features of the three software programs.

The preliminary analysis during site selection allows Archive staff to note site characteristics which may influence the choice of software and settings. It further identifies potential problems which may require alternative capture methods and/or manual collection and editing after the automated capture is complete. Section 5 summarizes common technical problems encountered.

3. CAPTURE

Actual capture of the websites takes place as close as possible to election day. Retaining the "look-and-feel" of the original websites is the primary goal of the collection, second only to retaining the content. This requires a great deal of manual intervention on the part of limited staff so the decision was made at the onset to distinguish between website *capture*, in which the sites are acquired once at a specific point in time, versus web *crawling*, where the same sites are acquired periodically over time.

If there are few sites to capture the capture process may occur entirely on election day. For the more typical large elections the process begins up to a week before. Staff generally start with sites for minor races, saving the major races and ballot propositions for last as these are more likely to be updated at the last minute. If time permits, on election day Archive staff will review the sites captured the previous week and check to see if there were substantial changes. If so, those sites are re-captured.

Archive staff use the information from the preliminary analysis to decide which capture software and settings to use for the first attempt. Once the capture is complete, staff then perform a quick check of the captured content to make sure the process worked properly. The preliminary analysis serves as a guide to probable errors. If errors are found the capture process may be repeated with different options or entirely different software. Archive staff may try four or five different capture methods before deciding that the whole site cannot be captured, in which case the home page alone is captured. Throughout the process staff make notes about the various capture attempts, both for later inclusion in the administrative metadata and to indicate errors which can be fixed by manual download of missing files or editing of broken links.

4. WEB CAPTURE SOFTWARE

The Archive uses three programs to capture and convert websites for archival storage. In addition, the capability of standard internet browsers (such as Internet Explorer and Firefox) to capture individual pages and their embedded content is used as a fallback method when none of the three programs is able to adequately capture an entire site.

Each of these programs has two primary functions. First is the actual capture process. Starting with a specific seed page, the software downloads and stores the html file along with all of the embedded files (images, media, style sheets, JavaScripts, Flash, PDFs, Word or Excel documents, PowerPoint presentations, etc.) which make up that page. From that seed page the software then follows hyperlinks to crawl the entire site, downloading all linked pages and their embedded files. This process continues until the entire site is captured or pre-defined limits are met.

The second function, and the more difficult, is to track and maintain all of the links that connect the individual pages and the links between html files and embedded files. This process can be extremely complex and may involve re-naming files, editing links, and re-organization of directory structures. Most notably, all absolute links must be converted to relative links, as the archived site will not forever reside in a specific directory of a specified domain. This function becomes exponentially more difficult as the size of the website increases. Furthermore, the more file and directory re-naming which takes place the more the archived site loses meaningful filenames which simplify later editing and which may be considered an aspect of the site worthy of preserving.

The following sub-sections present short summaries of the strengths and weaknesses of the three programs used by the Archive. (This evaluation applies to the versions listed. Characteristics may change in other versions.) Some of these issues are elaborated upon in the description of common technical problems in section 5.

4.1 WebCopier 4.4

- commercial software
- http://www.maximumsoft.com
- simplest and easiest to use of the three
- produces the most straightforward output, which is easiest to edit later
- tends to crash on large sites
- least sophisticated in its ability to parse css and js files
- generally the least powerful at handling complicated sites
- some problems translating non-ASCII characters

4.2 HT Track Website Copier 3.33

- free open source software
- http://www.httrack.com
- best program of the three at handling exceptionally large websites
- has powerful filtering tools and numerous options to control file re-naming and directory re-structuring
- highly technical, non-intuitive interface may be difficult for non-specialists
- pages at the maximum depth limit do not convert links that go back to higher levels
- slowest of the three

4.3 Offline Explorer Pro 4.6

- commercial software
- http://www.metaproducts.com
- fastest
- best at handling sites on multiple domains
- allows separate control of maximum capture depth for internal and external links
- only one which can handle some (but not all) Flash links

- does not re-name files to use htm or html extensions (This not only makes editing more difficult but also may threaten the long-term accessibility of the content, as future browsers may not recognize the files as html.)

5. COMMON TECHNICAL PROBLEMS

While every website presents its own unique challenges, there are a number of technical problems which are common to all three capture programs. Archive staff look for these problems when assessing and capturing sites as they may require extensive manual editing during the quality control process.

5.1 Dynamic Server-side Technologies

It is increasingly common for even small sites to use server-side technologies such as Active Server Pages, Cold Fusion, or PHP to produce dynamically generated content. Despite the dynamic processes used to manage the site, the actual files delivered to the user's browser are regular html files. Furthermore, most such sites still contain only a finite number of pages. Because of this the web capture programs can still crawl these sites and convert a dynamic site into a static snapshot. Nevertheless, some sites do have nigh-infinite content such as dynamically generated calendars or other "crawler traps". This is usually avoided by placing depth or level limits, but this can be a crude cutoff which excludes desired content in order to avoid the trap. Even with depth limits dynamic sites can be a challenge for the capture software as they may require extensive re-naming of files.

5.2 Links in non-HTML Files

Cascading Style Sheet (css), JavaScript (js), and Flash (swf) files provide a common problem: each is a file typically embedded into an html page, but unlike image or media files they can contain links to other files, including to other css, js, or swf files—which may then link to tertiary files, and so on. Each file format presents its own challenges.

Cascading style sheets and JavaScript files are both text files which can be directly parsed by the capture software—if the capture software is sophisticated enough to do so.

JavaScript files present a further problem due to the unique characteristic that any links contained within the js file are considered to originate from the location of the html file which called the js file, not from the js file's location. This can lead to complicated issues of directory addressing which most sites get around by using absolute addresses. Since that is not an option for the archived copy of the site, in such situations the archived site must (a) have multiple versions of the js file, one for each possible directory level, (b) duplicate the files linked from the js file in multiple locations, or (c) flatten the site's directory structure, a capture option which requires extensive re-naming of files and links. Though (a) and (b) are preferred solutions, none of the three capture software programs used by the Archive will do those automatically, necessitating manual editing to produce such results.

Flash files are unique in that they are a complex proprietary software format. Only one of the programs (Offline Explorer Pro 4.6) used by the archive is able to crawl links inside Flash files, and even that one cannot edit those links when it's necessary to correct for absolute links or variant directory structures.

5.3 Multiple Domains

One of the key issues in capturing a website is defining the boundaries of the "site". The nature of the web is such that an automated crawler could potentially continue following links forever. Luckily most sites can be defined by the domain on which they reside, and the default setting for all of the capture software programs is to limit the crawl to a single domain, ignoring external links. Typical capture options allow exceptions to this limit for images, media, and other embedded files, which often reside on external servers. However, as web technologies progress it is increasingly common for sites to integrate content from multiple domains. This also happens accidentally when sites use absolute addressing to their domain both with and without the www prefix. This situation presents a serious challenge to the capture software, and while all have some ability to handle multiple domains, none have a solution that works in all cases.

5.4 Streaming Media

Streaming media employs server-side web technologies to provide video or audio in a continuous stream rather than in whole files. While this allows faster access by the user, who does not have to wait for the entire file to download before viewing, it does not lend itself to archiving. In most cases the location of the original, non-streaming media file being streamed can be found within the code. In these cases the Archive downloads the original media file and edits all links to point directly to it, bypassing the streaming file, which is inactive without the proper server technologies. None of the three programs have this feature, so these changes must be made by hand.

6. QUALITY CONTROL

It is the nature of election sites that they become outdated and superfluous immediately after the results are announced. Many site owners abandon the site for days, weeks, or months, leaving it unchanged or simply adding a congratulatory or consolatory note to the home page. However, a significant number quickly delete their sites or re-purpose them to serve other functions. Because of this, detailed quality checking and editing must be completed as quickly as possible after the election, since the process may entail examination of the original site or downloading missed files, actions which become impossible as soon as the site owners make substantial changes.

The detailed review of the website involves examining every page, or at least a sufficiently large sample, to ensure that all pages and embedded files were downloaded and that the links between them are functional. This is done by skimming through the site in a browser, comparing the archived content with the original on the web, searching and scanning the html files to ensure that links have been translated properly and are not pointing back to the original site, and confirming the existence of files in their proper directories. Missing files are identified and downloaded from the original site, and mis-translated links are corrected using standard html editing software. (Links which were broken on the original site, along with any other coding errors, are not corrected.) In larger more complex sites, this process may require extensive analysis of the coding and interaction between style sheets and JavaScripts, the use of detailed global search and replace functions, and sometimes laborious editing of hundreds of individual links by hand. In addition to html editing software (the Archive uses HomeSite 4.5), the Archive has also found it necessary to purchase a Flash editing program, URL Action

Editor 5.11. This program allows the editing of links inside a compressed Shockwave Flash (swf) object downloaded from the web without access to the original Flash (fla) file from which it was derived, even though swf files were never intended to be editable.

The final goal of this process is to create an archived version of the site which maintains not only the content but also as much as possible of the look, feel, and browsing functionality of the original. Forms that rely on server-side technology (usually those for communicating with the campaign) are visible on the archived pages but are typically non-functional. Interactive features (such as Flash) are archived where possible. Links to files or sites outside the archive are left intact, but no attempt is made to maintain these links or to indicate to archive users that by following the link they are leaving the archive.

7. METADATA

Once the websites have been reviewed and edited, metadata is manually created and input into a Microsoft Access 2003 database. The descriptive metadata elements (title, date, subject, description, language, type, format, coverage) are based on Dublin Core and reflect an attempt to balance the anticipated searching needs of end users with the costs associated with the creation of metadata. Subjects are derived from a combination of Library of Congress subject headings and locally defined authority lists developed for the original print Campaign Literature Collection.

Administrative metadata is derived from the detailed notes created by staff during the capture and review process. Every record contains the date of capture, URL of the original website and the capture software(s) used to archive and edit the site. When applicable any substantial modifications made to the site are also documented.

8. ACCESS

The UCLA Digital Library Program (DLP) provides centralized access to the Library's digital collections. The UCLA Core database schema 3.2 is represented in both a Microsoft Access database and an Oracle 9i relational database system. Microsoft Access acts as a data collection tool as well as a data filter. A data transfer Java program is run which connects to Access and Oracle and transfers the data. Once the Access database and archived websites are uploaded, the collection is accessed via the DLP's virtual collection system which runs under Adobe (formerly Macromedia) JRun Application Server. The database server (Oracle) runs under the Linux operating system. The application server (Java) runs under the Windows XP operating system.

The virtual collection system supports browsing, keyword and phrase searching, and truncation and wildcards. Keyword searching is available from the basic search screen, while the advanced search screen allows users to search for term(s) within specific fields (keyword, title, subject, language, type), limit by date, and sort results. Users also have the option to browse the collection by subject or view all archived websites associated with a specific election. Initial search results are returned in a brief format which includes links to the full metadata record and the archived website. Results can be added to customized "virtual collections" which allow users to store and add notes to the individual records. Virtual collections saved in the public mode support collaboration by allowing multiple users to add new records and to edit and add notes to existing records.

When originally created, the Archive's content resided within the web pages of the UCLA Young Research Library and was accessed via a series of browsable index pages. This was not considered an ideal solution, as the index listings required extensive work to create and there was no method of searching the Archive's metadata. To address this issue, in September 2004 the Archive's contents were copied into the UCLA Library's Digital Library Program, where the much more sophisticated search and browse interface described above was available. Subsequently, in November 2004 the browsable index pages were taken down.

Unfortunately, though the new interface provided many access advantages, by removing direct links accessible to web crawlers the Archive was inadvertently removed from the public web and placed into the deep web. This precipitated a dramatic drop in the Archive's usage as the contents ceased to be discoverable on Google or other web search engines. In 2004 the Archive received 90,319 visits. In 2005 that number dropped to 6,897, and in 2006 there were 9,210 visitors. (Number of visitors was calculated from server logs using WebTrends Analytics 8.)

Archive staff are still working on how this issue should be addressed. The simplest solution would be to create a single page full of links to all of the individual archived websites. If that page itself were linked from anywhere within the Library's web content, the full contents of the archive would once more become visible to crawlers.

9. COPYRIGHT

In keeping with the UCLA Library's mission to provide access to information resources in support of the research and instructional mission of the University, the archived websites are made available based on the principles of fair use combined with an opt-out policy. Contact information found on the original website or gleaned from other sources (such as the Secretary of State's Certified List of Candidates) is recorded and used to notify the copyright owners that their website has been captured and will be made accessible via the UCLA Online Campaign Literature Archive interface. Copyright owners are given the option to opt-out of the Archive and request that the captured website or specific portions of it be removed from the collection. The archived content is completely removed from the collection as opposed to retaining it in a dark archive. The Archive honors the right to remove specific pages or files, but will not alter content within a page or file. In nine years, only one take down request has been received (and honored).

10. PRESERVATION

The nature of archived websites defies some of the standard definitions used for image or document preservation. There are no "master files"; by their very nature, the archived sites are derivatives, since the capture process itself involves extensive transformation and editing of the files. However, they remain masters in the sense that all files are stored in their native, uncompressed, web-compatible formats organized in their original directory hierarchy. This is due to the necessity of retaining the *relationships* between the files, relationships which turn thousands of individual files into a unified website. This mechanism relies on the end user having access to a contemporary web browser in order to re-assemble the files in the manner they were originally intended to be viewed. This is not ideal from a preservation point of view. The Archive is relying on the ability of future users to possess or emulate an early 21st century browser. This is not an

unreasonable assumption, given that the web environment is based on a suite of widely accepted, relatively simple, and well-documented protocols. Archive staff are keeping abreast of developments in alternate preservation methods and may explore these possibilities at some point.

"The Digital Library provides backup and digital preservation support for all its collections. Digital master files are stored in their original, uncompressed format and can be made web accessible. Backup services are provided by Library Information Technology. In addition, the Library is working with the CDL [California Digital Library] Digital Preservation Program to provide for long-term preservation of the UCLA Library's digital assets." [1] The Archive's collection is backed up to tape each night and stored offsite. Eventually it is hoped that copies of the archived websites will be deposited into the CDL's digital preservation repository which provides long term secure storage, object integrity, preservation, and format migration strategies for the University of California libraries.

11. CONCLUSION

Nine years have passed since the Archive's staff first began to capture local and state campaign websites. Researchers are beginning to see the inherent value possessed by this primary content. "Online official materials, campaign materials, voter guides, and some news items for past elections about two California offices and one California proposition were obtained from the UCLA Online Campaign Literature Archive." [9] "UCLA's Online Campaign Literature Archive…was also very helpful in this regard." [8] Links to the Archive are showing up on university course web pages [3], library subject guides [2, 5], and within Wikipedia articles [4].

As elections at all levels increasingly move online, the ability to capture and preserve the digital artifacts of political campaigns will be critical to scholars. Foot and Schneider "quickly decided [they] would need to begin archiving election-related Web sites" [6] as they studied the use of the web in political campaigns. They created a digital supplement to their printed work which links to the archived websites being discussed in the text. In the future researchers interested in examining primary source materials will need to rely on archived websites.

The authors hope that sharing their expertise and experience with the broader web archiving and library community will aid those involved in the development of web archiving projects as we all race to preserve the online record of our time.

12. REFERENCES

[1] About the Digital Library Program. http://www2.library.ucla.edu/libraries/2632.cfm

[2] Election Information. http://library.csun.edu/Find_Resources/Government_Publications/election.html

[3] The History of American Presidential Campaigning: Readings. http://www.arts.mcgill.ca/history/faculty/TROYWEB/Courseweb/hist301readings.htm

[4] Steve Soboroff: Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Steve_Soboroff

[5] Woodbury University Library: Political Science. http://web3.woodbury.edu/library/resources/polisci.html

[6] Foot, K.A. and Schneider, S.M. *Web campaigning*. MIT Press, Cambridge, Mass., 2006.

[7] Masanès, J. Web archiving methods and approaches: A comparative study. *Library Trends*, *54* (1): 72-90, 2005.

[8] Masket, S. If it Walks Like a Party... The Emergence of Unofficial Party Organizations in California *Prepared for Delivery at the Third Annual Conference on State Politics and Policy*, Tucson, Arizona, 2003.

[9] Robertson, S.P., Achananuparp, P., James, L.G., Park, S.J., Zhou, N. and Clare, M.J. Voting and political information gathering on paper and online *CHI '05 extended abstracts on Human factors in computing systems*, ACM Press, Portland, OR, USA, 2005.