Political Communications Web Archiving: Addressing Typology and Timing for Selection, Preservation and Access

Bernard Reilly[1], Carolyn Palaima[2], Kent Norsworthy[2], Leslie Myrick[3], Gretchen Tuchel[1] and James Simon[1]

[1]Center for Research Libraries, 6050 South Kenwood Avenue, Chicago IL 60637
(reilly@crl.edu, tuchel@crl.edu, simon@crl.edu) http://www.crl.edu
[2]Latin American Network Information Center – LANIC, Teresa Lozano Long Institute of Latin American Studies, 1 University Station D0800, Austin, Texas 78712
(c.palaima@mail.utexas.edu, kent@lanic.utexas.edu) http://lanic.utexas.edu
[3]Digital Library Group, New York University Libraries, Bobst, 70 Wash Sq S, 807, New York, NY 10012
(leslie.myrick@nyu.edu) http://www.nyu.edu/library/bobst/collections/digilib/

Abstract. The Center for Research Libraries has undertaken a research and planning grant to help ensure long-term availability of important documents and messages disseminated via the World Wide Web by non-governmental political groups and parties. As an introduction to the project, this paper focuses on the curatorial challenges of selection, typology, and timing for archiving the very broad and disparate array of political communications sites. In preserving political Web communications, defining typology assists curators in their selection and identification process, in creating controlled metadata vocabulary and in creating a subject-driven access portal. Timing studies assist in determining frequency of capture for ephemeral Web material.

1 Introduction: Project, Participants, Goals

The Political Communications Web Archiving Project is a research and planning initiative under the coordination of the Center for Research Libraries (CRL) and funded by a grant from the Andrew W. Mellon Foundation. The investigation seeks to help ensure long-term availability of the important documents and messages disseminated via the World Wide Web by non-governmental political groups and parties. These kinds of materials comprise a valuable source of historical and social science information but they are by nature fugitive and susceptible to loss. This project will lay the methodological groundwork for their cooperative preservation. The joint planning effort focuses on four world regions, each under the responsibility of the Project's four university partners: Cornell University (Southeast Asia), New York University (Western Europe), Stanford University (Sub-Saharan Africa), and the University of Texas at Austin (Latin America). Also participating in the effort are the San Francisco-based Internet Archive and the Library of Congress.

Within the past decade the World Wide Web has emerged as a vital medium of political communication. It now serves political activists, parties, popular fronts, and other non-governmental organizations (NGOs) as a global message board through which to communicate with constituents and the world community. The Web provides a widely accessible and relatively unrestricted medium for rapid broadcast of information and public posting of critical documents such as manifestoes, constitutions, declarations, and treaties.

Such communications are the digital-era counterparts of the posters, pamphlets, broadsides and other forms of street literature that have long served as an indispensable source of information on political ideology and strategy. Unfortunately, the content of Web-based political communications is disappearing without being properly archived. While extremely important for scholarly research in the humanities and social sciences political communications have received little attention in national efforts to devise a preservation strategy for digital materials. Those efforts have tended to focus on materials such as electronic journals, moving image, audio, and commercial broadcast media, which are generated by commercial producers and organizations and where preservation strategies can involve the cooperation of the producers. Political communications in print and electronic form are more ephemeral precisely because of the inherent difficulty in engaging the producers, usually political activists with no long-term interest in the archiving activity.[1]

The CRL investigation is using Web communications produced by political groups in Southeast Asia, Latin America, and Sub-Saharan Africa and by radical organizations in Europe as a test bed of materials. The effort is in the process of elaborating a framework, methodologies, and technical specifications for three aspects of ongoing, sustainable archiving:

- *Long-term resource management*: The organizational and economic framework necessary to support archiving, management, and preservation of Web political materials on an ongoing, cooperative basis.
- *Curatorship*: The optimal curatorial regimes and practices for identification, targeting and capture of Web political communications to be archived.
- *Technology*: The general technical requirements, specifications, and tools best suited to the capture and archiving of political communications.

The project builds upon the investigations currently underway at the partner universities, the Internet Archive, and the Library of Congress, and will draw conclusions and identify methodologies that can be applied to the harvesting of similar materials from all regions.

The primary uses of the archived materials are:

- Scholarly research and teaching, in particular by historians and political scientists;
- Study and informational use, by members of the international development, policy, diplomatic, and journalism communities, and lay individuals;
- Inclusion in non-commercial publications/aggregations (TBD).

This paper does not attempt to cover the extensive research and evaluation being conducted for all three aspects of the project listed above. As an introduction to the

project, this paper focuses on the curatorial challenges of selection, typology, and timing for archiving the very broad and disparate array of political communications sites. Although project participants have reached broad agreement on copyright and intellectual property rights pertaining to the archive, these topics are beyond the scope of this paper.

2 Identification and Selection

Content eligible for inclusion in the Archive includes static Web sites and documents in all formats mounted on the surface Web. "Sites" here refers to a collection of interlinked Web pages, including a host page, residing at the same network location. These pages may include HTML files and all embedded or linked documents and files, including text, image, sound, and moving image files. Also to be considered for inclusion in the Archive, subject to further discussion by the curatorial team, are:

- newsgroup postings (such as the Usenet postings archived by Google, cf. http://www.google.com/googlegroups/archive_announce_20.html)
- Web site "defacements" (cf. "Analysis of the Defacement of Indian Web sites" *First Monday* http://www.firstmonday.dk/issues/issue7_12/srijith/index.html and http://www.civilservices.gov.in/lbsnaa/index.jsp#)

Related materials under the political communications rubric that might be addressed by subsequent investigations include deep Web materials that are password-protected or otherwise designed to be resistant to robots and access; listserve digests; RSS feeds; and databases.[2]

Political communications Web sites in developing regions bring their own complexities to the identification process. Multiple and unrelated sites are often hosted on a single domain name, many sites are event driven, levels of political instability may result in site volatility, and due to issues of connectivity and repression, sites may be maintained out of country. A consensus was formed early in the project that curators and area specialists would play a major role in the selection process by providing seed URLs based on a collection policy statement and guidelines. These would be added to a list of sites for scheduled automated crawls. And, as event-driven sites emerge, curators would identify them for timely capture. The need to be at the same time comprehensive was recognized and the ideal curatorial regime for targeting materials will combine manual and automated operations. Accordingly, one possible harvesting model discussed combines three elements: selection by curators; wholesale web crawling with smart, or "learning" robots;[3] and reliance on larger, comprehensive Web archiving projects. Such a model might involve human and automated as well as centralized and distributed processes:

1. Curatorial / selection activities – identification of the characteristics (e.g., URL, domain, presence of keywords, etc.) of desired sites; specification of capture regime (periodicity, extent, etc.); negotiation of permissions.
2. Central/automated functions - utilities and services include: operates utilities, programs crawl, harvests, archives, aggregates; provides certification/

documentation of content authenticity; supplies feedback on results/data from target sites, e.g., occurrence of new links, trends, resistance, patterns, etc., to inform selection.

3. Backup functions: IA / Google / PANDORA / other organization harvests on a wholesale basis as backup to selection; IA makes materials available for retrospective harvesting.

3 Issues in Defining Typology

The project's Curatorial Team is working on the elaboration of a classification scheme with a controlled vocabulary for the types of sites that will be included in the Archive. This scheme will be a permanent fixture of the project in at least three distinct areas:

- As a tool used by curators in the selection and identification phase of collection activities;
- As a controlled vocabulary for populating descriptive metadata elements, such as "subject" and "keywords"; and,
- As part of an end-user, directory-driven subject access portal to the archived sites.

Several key issues regarding such a scheme or typology have been identified to date. First and foremost is the recognition that, given the widely divergent nature of politics and the "political process" in these four regions of the world, a narrowly defined typology will create numerous problems. The current proposal is to thus elaborate a broad and exhaustive scheme which would likely wind up including numerous categories and sub-categories that might be specific to just one region.

Currently, the Curatorial Team is evaluating the categories pertaining to politics contained in the UNESCO Thesaurus, most of which correspond to categories of political actors or groups and doctrines, which would be augmented by region-specific categories submitted by curators representing the four regions covered by the project, as necessary.

Preliminary research has confirmed that a significant portion of the Web sites that would be included in the Archive are "event-focused" as opposed to "actor-focused." Thus, another approach under examination is to essentially employ a dual classification scheme, one that would comprise categories for type of political group (political party, armed insurgent, NGO, etc.) as well as categories based on the nature of the event around which the site is focused (election, plebiscite, coup d'etat, insurgency, etc.). The UNESCO Thesaurus, to a certain extent, exhibits this quality, but here again, it would need to be enhanced with the addition of region-specific types of events.

For robust search and retrieval by end users of the archive a combination of automatically generated and manually tagged metadata will be required to cover the three categories of metadata generally associated with digital objects: descriptive, structural, and administrative; the last being some combination of technical, rights, source, provenance and preservation metadata. A metadata system that integrates all three categories into an extensible, flexible package is the end goal.[4] Project members

are evaluating harvesters to review quality of site capture and automatically generated metadata.

4 Timing Issues

The primary issue here has to do with "frequency of capture," which is of crucial importance for two reasons:

- The average life span of a page on the Web is estimated to be just 44 days.[5] Sites that are identified for capture and inclusion in the Archive will have to be crawled in a timely manner or they may be lost forever.
- Content updates on some target sites will take place numerous times during a single day, while other sites remain unchanged for years.

Preliminary research on the types of sites that would be included in this Archive has revealed that the frequency of content updates on the sites varies widely depending on numerous factors, including the nature of the group itself or of the event with which the site is connected.

Project staff conducted two exercises to identify and document timing issues in relation to Web coverage of electoral processes. One exercise, conducted in late 2002, involved sites from the *LANIC Electoral Observatory*, a directory of links to Latin American election sites. A total of 148 URLs from the *Observatory* were searched through the Internet Archive's (IA) *Wayback Machine*. These URLs corresponded to a total of 21 separate electoral processes held in 15 countries of Latin America between December 1998 and June 2002.

A full 61% of the 148 sites had already disappeared from the Web by the time the exercise was conducted, thus reinforcing this Project's rationale regarding the volatility and fugitive nature of political communications on the Web. Of the 148 total sites, 19 sites (13%) were not present in the Internet Archive at all. Of the 129 sites with a presence in the Internet Archive, in 36% of the cases, the crawled versions of the site available in the Archive did not include the critical time period leading up to the elections themselves. This highlights the need for curator-driven input into crawl scheduling and frequency of capture in the case of time-sensitive political events.

In spring of 2003, a separate exercise was conducted looking at 38 Internet Archive-crawled sites related to the Nigerian electoral process. Results similar to the Latin American exercise were obtained, reinforcing the need for the archiving project to build in mechanisms that will facilitate systematic curator input on the capture schedule for selected sites.

Based on these findings we propose a two-tiered approach, automated and selective, whereby one set of curator-identified sites would be crawled according to a periodic schedule (i.e., Group A sites weekly, Group B sites monthly, Group C sites quarterly, etc.), while another set of sites would be identified by curators on the basis of current or "breaking" events and fed into the crawl along with a frequency timetable (i.e. "four times per day for the next two weeks, daily for two months, etc.).

5 Acceptable Levels of Loss

No combination of curatorial regime and technological platform will ever succeed in capturing "all of the content all of the time" that a researcher of Web-based political communications might want. There will always be some degree of content "loss" from what existed on the Web live at a point in time, and what is retrievable from the Archive.

As part of the LANIC Electoral Observatory exercise described above, we also documented "levels of loss" for the 129 electoral sites that were present in the Internet Archive. For 12 of the 129 sites (9%), there was essentially no access to the site content at all even though the site appeared in the IA's index as having been captured. In some of these cases the site Home Page displayed, but clicking on any of the content sections brought up an error message, in others requesting the Home Page itself resulted in an error message.

In 78 of the sites, or 60% of the 129 in the Archive, there was a significant loss of content, such as image files missing, an inability to access content beyond the second or third level of the site, or non-functioning site navigational aids. Typically, we had better success retrieving adequate content from the Archive of sites relying on older Web techniques such as static HTML pages, and less success with sites that tended to incorporate newer or more advanced technologies or that were generally more complex from a design or information architecture standpoint.

Project staff are in the process of planning a series of surveys and exercises to be conducted with scholars and area specialists to ascertain what would be an "acceptable level of loss," that is, defining a point beyond which a capture regime would result in a product which would essentially not be of sufficient use to a researcher.

6 Related Works

Political Web Archiving/Portals

Archipol Netherlands
http://www.archipol.nl

IPOP
http://www.ipop.org.uk

LANIC
http://lanic.utexas.edu

Occasio: Digital Social History Archive
http://www.iisg.nl/occasio/index.html

SOSIG
http://sosig.esrc.bristol.ac.uk/roads/subject-listing/World-cat/polcom.html

The Welsh Political Archive
http://www.llgc.org.uk/lc/awg_s_awg.htm

Political Web Use Studies

Institute for Policy, Democracy and the Internet
http://democracyonline.org

National Election Studies
http://www.umich.edu/~nes

Politicalweb.info
http://politicalweb.info/home.html

Stanford University Political Communication Lab
http://pcl.stanford.edu

General Digital Preservation

Beagrie, Neal and Greenstein, Dan: A Strategic Policy Framework for Creating and Preserving Digital Collections
http://www.ahds.ac.uk/old/manage/framework.htm

CLIR & DAI: The State of Digital Preservation: An International Perspective
http://www.clir.org/pubs/reports/pub107/pub107.pdf

Hodge, Gail: Best Practices: An Information Life Cycle Approach
http://www.dlib.org/dlib/january00/01hodge.html

Holdsworth, David, Wheatley, Paul: Emulation, Preservation, and Abstraction
http://www.rlg.ac.uk/preserv/diginews/diginews5-4.html#feature2

Kenney, Ann R., Nancy Y., Botticelli, Peter, Entlich, Richard, Lagoze, Carl, Payette, Sandra.: Preservation Risk Management for Web Resources: Virtual Remote Control in Cornell's Project Prism
http://www.dlib.org/dlib/january02/kenney/01kenney.html

Lavoie, Brian: The Incentive to Preserve Digital Materials: Roles, Scenarios and Economic Decision-making
http://www.digicult.info/downloads/digicult_info3.pdf

Lavoie, Brian: Meeting the challenges of digital preservation: The OAIS reference model http://www.oclc.org/research/publications/newsletter/repubs/lavoie243/

Lupovici, Catherine, Masanès, Julien: Metadata for long term-preservation
http://www.kb.nl/coop/nedlib/results/D4.2/D4.2.htm

Moore, Regan, Baru, Chaitan, Rajasekar, Arcot, Ludaescher, Bertram, Marciano, Richard, Wan, Michael, Schroeder, Wayne, Gupta, Amarnath : Collection-Based Persistent Digital Archives - Part 1
http://www.dlib.org/dlib/march00/moore/03moore-pt1.html

Moore, Regan, Baru, Chaitan, Rajasekar, Arcot, Ludaescher, Bertram, Marciano, Richard, Wan, Michael, Schroeder, Wayne, Gupta, Amarnath: Collection-Based Persistent Digital Archives - Part 2
http://www.dlib.org/dlib/april00/moore/04moore-pt2.html

RLG Final Report: Preserving Digital Information
http://www.rlg.org/ArchTF/

RLG/OCLC: Trusted Digital Repositories
http://www.rlg.org/longterm/repositories.pdf

RLG/OCLC: Preservation Metadata and the OAIS Information Model
http://www.oclc.org/research/pmwg/pm_framework.pdf

Russell, Kelly, Ellis Weinberger, Ellis: Cost elements of digital preservation
http://www.leeds.ac.uk/cedars/colman/CIW01r.html

Wheatley, Paul: Migration - a CAMiLEON discussion paper
http://www.ariadne.ac.uk/issue29/camileon/

Web Archiving

Beagrie, Neal, Pothen, Philip, eds.: Web-archiving: Managing and Archiving Online Documents and Records -- Ariadne Issue 32
http://www.ariadne.ac.uk/issue32/web-archiving/

Bergman, Michael K.: The Deep Web: Surfacing Hidden Value
http://www.press.umich.edu/jep/07-01/bergman.html

Boudrez, Filip, von den Eynde, Sofie: Archiving Websites
http://www.antwerpen.be/david/teksten/Report5.pdf

Royal Library of Denmark, Archiving Web Publications
http://www.kb.nl/kb/ict/dea/ltp/reports/6-webpublications.pdf

Lyman, Peter: Archiving the World Wide Web
http://www.clir.org/pubs/reports/pub106/web.html

Political Web Archiving

Craig, Ann: Bridging the Digital Divide: State Government as Content Provider. The Illinois Experience http://www.firstmonday.dk/issues/issue6_4/craig/index.html

Gupta, Amaranth: Preserving Presidential Library Websites
http://www.sdsc.edu/TR/TR-2001-03.pdf

Peach, Martha: Archiving the Internet, Web Pages of Political Parties (1998)
http://www.sosig.ac.uk/iriss/papers/paper24.htm

Metadata Extraction

Adams, Katherine C.: Extracting Knowledge
http://www.intelligentkm.com/feature/010507/feat1.shtml

DC-DOT RDF extractor
http://www.ukoln.ac.uk/metadata/dcdot/

Liddle, Stephen W.: On the Automatic Extraction of Data from the Hidden Web
http://www.deg.byu.edu/papers/daswis01.pdf

Pierre, John M.: Practical Issues for Automated Categorization of Web Sites
http://www.ics.forth.gr/isl/SemWeb/proceedings/session3-3/html_version/semanticweb.html

---

[1] Project PRISM is developing a risk management methodology for the preservation of web sites: Kenney, Ann R., Nancy Y., Botticelli, Peter, Entlich, Richard, Lagoze, Carl, Payette, Sandra.: Preservation Risk Management for Web Resources: Virtual Remote Control in Cornell's Project Prism (http://www.dlib.org/dlib/january02/kenney/01kenney.html).

[2] The BnF has been exploring the feasibility of harvesting the deep web: Masanes, Julien: Towards Continuous Web Archiving: First Results and an Agenda for the Future: (http://www.dlib.org/dlib/december02/masanes/12masanes.html. See also Bergman, Michael K.: The Deep Web: Surfacing Hidden Value http://www.press.umich.edu/jep/07-01/bergman.html; Raghavan, S., Garcia-Molina, H.: Crawling the hidden Web. http://www.dia.uniroma3.it/~vldbproc/017_129.pdf.

[3] One such smart crawler is Mercator (http://research.compaq.com/SRC/mercator/). Heydon, Allan, Najork, Marc: Mercator, a Scalable, Extensible Web Crawler (http://www.research.compaq.com/SRC/mercator/papers/www/paper.html). The BnF has been using the Xyleme crawler (http://www.xyleme.com/en/index_html).

[4] The prime candidate is the METS standard (http://www.loc.gov/standards/mets/), which has been developed to serve as a SIP, an AIP and a DIP in an OAIS-compliant system. The technical team will be evaluating the feasibility of using METS for metadata management, access and object navigation.

[5] Lyman, Peter: Archiving the World Wide Web: (http://www.clir.org/pubs/reports/pub106/web.html).