Political Communications Web Archiving
An Investigation Funded by the Andrew W. Mellon Foundation

Center for Research Libraries
Latin American Network Information Center, University of Texas at Austin
New York University
Cornell University
Stanford University
Internet Archive

June2004

Contents

Appendices

Curatorial Investigation
1. Archival Access Policy survey
2. LANIC Electoral Observatory exercise results (IA Assessment)
3. Nigerian Election 2003 Crawl – Curatorial Assessment
4. Timing Exercise (HTTrack Assessment)
5. Typology of sites (UNESCO Thesaurus)
6. Test Data Input Module (MODS descriptive Data)
7. Lor, Peter and Britz, Hannes: "A South-North Perspective on Web Archiving," November 8, 2002.

Technical Investigation - Topical Reports
8. Technical Challenges of Web Archiving
9. Digital Preservation Considerations for Web Archiving
10. Risk Management for Web Resources
11. Web Archiving Cost Issues
12. Summary of Staffing Requirements for Ten Web Archiving Projects

Technical Investigation - Evaluations of Prototypes
13. Comparative Merits of Current Methodologies
14. Longer Evaluation of PANDORA/Kulturarw3
15. Evaluation of WARP

Technical Investigation - Evaluations of Harvesters
16. Harvester Evaluation
17. Case Study: NEDLIB Harvester
18. Case Study: PANDAS/HTTrack
19. Summary of Mercator crawl problems

Technical Investigation - Metadata; OAIS; METS and Websites
20. The Feasibility of Automatic Metadata Harvest from Crawler Logs
21. The Feasibility of Populating a METS File from an IA SIP (.arc + .dat)
22. Stripped-Down METS Template
23. .arcscraper output
24. .datscraper output
25. IA .arc Format and the OAIS Metadata Framework
26. IA .arc Format and OAIS -- Table
27. IA .arc Format and OAIS Preservation Description Metadata - Table

Technical Investigation - Crawler Reports on robots.txt and Meta Tag Usage
28. Robots.txt evaluation
29. Meta tag usage evaluation
30. Title Metadata from .arc/.dat files
31. The Case of the Purloined Metadata

Technical Investigation - Crawl Results
32. Summary of Nigerian Mercator crawls
33. HTTP Server Software Use: Nigerian Sites
34. Comparative page data for Nigerian Sites

Technical Investigation - Project Demos
35.   Oracle Intermedia Full Text Search Implementation
36.   Sample Archive Records – MODS Descriptors
37.   Sample Archive Records – METS Viewer

Long Term Resource Management Investigation
38.   Political Web Wire Frame document
39.   User Survey – political Web sites as primary source material
40.   User survey results - summary

Introduction: Aims and Background of the Investigation

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, statements, constitutions, declarations, and treaties. The use of information and communications technologies (ICTs) by political actors, particularly in the developing world, has emerged recently as an important field of study in the social sciences.

This report of the Political Communications Web Archiving investigation (PCWA) outlines broad strategies for archiving materials from the Political Web, to provide for the capture, preservation, and long-term availability of Web-based political communications for educational and research uses. If properly archived these materials will be valuable resources for historical studies and the social sciences. The potential benefits of preserving these materials, however, go beyond the academic community. The Political Web is a rich source of information and alternative viewpoints that can shape international relations and pubic policy, and overcome the increasing homogeneity of knowledge sources brought about by the growing consolidation of the commercial media world.

While the technical aspects of capturing and preserving Web sites present considerable challenges, the curatorial regimes and challenge of sustainability, which must necessarily inform the technical solutions, were a chief focus of the PCWA investigation. The matters of curatorship addressed included selection, timing and approaches to harvesting of Web communications, the "artifactual" characteristics to be preserved for archived Web content, and potential intellectual property constraints to be overcome.

Participants in the PCWA investigation also explored prospective strategies for maintaining the archiving activities over time, which ultimately will require creation of a framework for organizational and individual participation, underwritten by well-structured governance and diverse sources of funding. Investigators concluded that the form of governance best suited to support Political Web archiving activities with the highest likelihood of sustainability is a consortium model, controlled and governed mainly by the larger research community. The consortium would undertake the critical stewardship or "brokering" activities that the community requires. The consortium would also serve the non-academic research community, which has shown an ability to support dissemination and maintenance of traditional political materials.

The present report proposes a service model that is adaptable to and accommodating of evolving digital and network technologies. The report also specifies an organizational framework that will best support both ongoing digital collection development and the long-term maintenance of the archived resources. The model indicates the general costs and requirements of sustaining the underlying activities and infrastructure, the characteristics and requirements of entities that might perform the various archiving activities; where the responsibilities for such activities are best situated and the ideal configuration of relationships and suggests partnerships needed to ensure that those responsibilities are fulfilled.

The stewardship activities described in the PCWA model are comparable to activities undertaken by the Center for Research Libraries in some of the Center's traditional area studies resource-building programs, including the Area Microform Projects (AMPs), International Coalition on Newspapers (ICON), and the Digital South Asia Library. PCWA governance and management will have to involve a broader community of participants and advisors than the Center has embraced in the past. To adequately support Political Web archiving the Center will have to extend its membership and constituency to organizations and audiences beyond the higher education academic community.

Background: The Project, Participants, Goals

The Political Communications Web Archiving Project was 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 joint planning effort focused 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 were the San Francisco-based Internet Archive and the Library of Congress.

The investigation used 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 project also built upon investigations underway at the partner universities, the Internet Archive, and the Library of Congress, and drew conclusions and identified methodologies applicable to the harvesting of similar materials from all regions.

The investigation was conducted by three teams: Long-Term Resource Management, Curatorial, and Technical. This report was written by Carolyn Palaima, Leslie Myrick, James Simon, and Bernard F. Reilly, with contributions by Nancy McGovern and Kent Norsworthy.

Participants and Advisors to the Project

Bernard F. Reilly, Center for Research Libraries (Project Director)

Curatorial Investigation Team

Carolyn Palaima, University of Texas/LANIC (Team Leader)

Karen Fung, Stanford University
Michael Nash, New York University
Kent Norsworthy, University of Texas/LANIC
Nancy McGovern, Cornell University
Allen Riedy, University of Hawaii (formerly at Cornell University)

Advisors:

Angel Batiste, Library of Congress
Carolyn T. Brown, Area Studies, Library of Congress
David W. McKee, Deputy Director, Information Resources Program, U.S. Department of State
Robert Latham, Director, Program on Information Technology and International Cooperation, Social Science Research Council
Georgia Harper, General Counsel, University of Texas System
Peter Lor, Director, National Library of South Africa
Andrew H. Lee, Tamiment Librarian, New York University
Kirsten A. Foot, Department of Communication, University of Washington and WebArchivist.org
Steve Schneider, SUNY
Pamela Graham, Columbia University
David Hirsch, University of California at Los Angeles
Deborah Jakubs, Duke University
Dan Hazen, Harvard University
Ali B. Ali-Dinar, African Studies Center, University of Pennsylvania

Long-Term Resource Management Team

James Simon, Center for Research Libraries (Team Leader)

Karen Fung, Stanford University
Michele Kimpton, Internet Archive
Leslie Myrick, New York University
Nancy McGovern, Cornell University
Carolyn Palaima, University of Texas/LANIC
Bernard F. Reilly, Center for Research Libraries

Advisors:

Martha Anderson, Office of Strategic Initiatives, Library of Congress
Daniel Greenstein, California Digital Library
Mary Summerfield, University of Chicago Press
Robert L. Worden, Federal Research Division, Library of Congress

Technical Investigation Team

Leslie Myrick, New York University (Team Leader)

William Kehoe, Cornell University/PRISM
Michele Kimpton, Internet Archive
Ning Lin, University of Texas/LANIC
Nancy McGovern, Cornell University/PRISM

Advisors:

Cassy Ammen, Minerva Project, Library of Congress
Patricia Cruse, California Digital Library
Richard Entlich, Cornell University
Abbie M. Grotke, Minerva Project, Library of Congress
Jerome McDonough, New York University
Igor Ranitovic, Internet Archive
Punya Rawal, Center for Research Libraries

1. The Political Web – Production and Producer Behaviors

Political Web content is generated and supported by:

1. Producers -non-governmental organizations, governments, political parties, and individuals engaged in political activities in various parts of the world.

2. Hosts – Some sites are hosted by the producing organization, government, or individual. But most producers make use of commercial or non-commercial ISPs.

For the most part, hosting behaviors for the Political Web are generally comparable to those of hosts for non-political materials, and have little impact on the content of the sites. They do, however, have a potential impact on the persistence of the sites, as shown by the risk management analysis in section four of this report. Host behaviors may also affect the ability of users and archivists to authenticate the sites. Chinese dissident groups, for instance, enlist the help of other sympathetic organizations and individuals who are willing configure their own computers to function as "proxy servers," enabling users in China to elude government efforts to block access to dissident Web sites.[1] And during the April 2003 elections in Nigeria many of the political parties maintained their Web sites abroad.

The use of information and communications technologies (ICTs) by political actors, particularly in the developing world, has emerged recently as an important field of study in the social sciences. Analyses of the Web-based publicity campaigns of the Zapatistas in Mexico, the attacks on East Timor government Web sites during the 1998 struggle, and the use of the Web by Islamic dissidents in the Middle East have been published in scholarly and public policy journals and monographs [2] PCWA investigators made use of this literature in their analysis of producer behaviors.

The investigation surveyed the range of types of Web sites and objects that make up the Political Web and examined the kinds of organizations and entities that produced them. The study focused on three distinct regions in the developing world: Latin America, Southeast Asia, and Sub-Saharan Africa. The test bed also included a topically defined subset of Web-based political communications from a fourth region as well. The topically defined subset focused on communications from radical groups in Western Europe.

Most of the producing organizations are formally constituted, stable entities. Some are legally incorporated. Others, like the Islamic Jihad in the Middle East and the FARC in Colombia, are loose affiliations or were formed on an ad hoc basis in response to particular events or political conditions. A significant number of producing organizations, however, either deliberately avoid disclosure of their membership, structure, and geographic location; or are so mercurial in their membership and operations as to defy definition as an entity.

Producer activities cover a complex array of endeavors. They include the creation of original political communications and information content, mounting of the content on the Web; arrangement for the hosting of that content on one or multiple servers; and support, revision and augmentation of the content and its functionality over time. In some instances producers also "archive" their own content in open or semi-open Web spaces. The Movement for Islamic Reform in Arabia (MIRA), for instance, maintains back issues of their electronic newsletters on their Web site, like many traditional news organizations. Some producers also mount Web sites that function as conduits for gathering subscribers for printed newsletters and newspapers (Muslim Brotherhood) or participants for authenticated online listserv discussions

---

[1] British Broadcasting Company "The World" National Public Radio, March 4, 2004 report on Chinese dissidents' circumvention of government-imposed Internet censorship to access foreign political news and communicate their message. Reference: http://www.theworld.org/latesteditions/20040304.shtml

[2] See: Harry Cleaver, "The Zapatistas and the International Circulation of Struggle" in John Holloway and Eloína Peláez, eds. *Zapatista! Reinventing Revolution in Mexico*, London:Pluto Press, 1998; Tedjabayu Basuki "Indonesia: The Net as a Weapon," *Cybersociology* 5:5, 1999; and Sean McLaughlin "The use of the Internet for political action by non-state dissident actors in the Middle East" First Monday, November 3, 2003. http://www.firstmonday.org/issues/issue8_11/

(MIRA).[3] Some NGOs, like BurmaNet, also operate notification services, which provide news to subscribing patrons and in some cases to others targeted by the producing organizations.

In general many static pages on the Political Web remain the same throughout their lifecycle. But the number of pages modified daily or weekly, e.g. dynamic pages generated from databases or RSS feeds, is significant.

The rate of change for political Web sites is difficult to generalize. Political sites, especially those maintained by radical groups and NGOs, are subject to bursts of activity around key events like elections, *coups d'etat*, and legislative debates. Many sites that come under the Political Web rubric essentially function as alternative news services, and undergo changes daily as events unfold. Similarly, the online output of some radical NGOs might replicate the ephemeral nature of street pamphlets and graffiti.

Production and maintenance of political Web sites is also affected by other factors, some of them less predictable, such as the financial or electoral fortunes of the producing entity, or government suppression of that entity. These factors affect not only the frequency of change, but the amount of content, sophistication of functionality, and reliability of site maintenance. The level and sophistication of functionality on political Web sites is also contingent upon the robustness of the technological infrastructure and environments in which the producers and their target audiences operate; and the power and reach of the regimes and other entities that the producing groups confront.[4]

The most important characteristic of Political Web materials, however, is their ephemeral character. In an attempt to quantify the fugitive nature of political communications on the Web, in October 2003 LANIC (Latin American Network Information Center) analyzed the rate of disappearance of sites on the LANIC Electoral Observatory, a directory of sites covering Latin American elections.[5] The Observatory chronicles elections in Latin America beginning with the Venezuelan election in 1998 through those occurring in 2003. The page of links for each election is not modified after the event is over. In reviewing the links to 226 sites, investigators found a steady attrition in the number of l sites that were still live after the elections, with an average of over 50% of all sites gone from the Web within two years.

Focused Web crawls of 37 sites mounted by Nigerian political parties and candidates surrounding the April 2003 Nigerian presidential and gubernatorial election yielded additional useful information about producer behaviors (See Appendices 3, 32-34). These crawls revealed a high rate of change in the target sites, many of which vanished within months of the elections. They also revealed the prevalence of out-of-country hosting of sites: 21 of the target sites were registered in the United States, five in Canada, five in the U.K.; and one each in Sweden and Albania. This suggests the limitations of a domain-specific approach to archiving which captures content relevant to a national domain.

The Nigerian sites, particularly political party sites, also employed a variety of applications, such as animated gif's, flash pages, and so forth. Studies of Web sites maintained by various political groups and interests in the Middle East, moreover, revealed that audio and video recordings, some quite lengthy, are integral to the "message" of those groups. Rafal Rohozinski, a specialist and senior advisor to the United Nations on conflict zones, writes in his study "Bullets to Bytes: Reflections of ICTs and 'Local' Conflict," that "activists and 'hacktivists' in the industrialized countries were quick to pick up on the potential of [digital video] technology, and in recent years independent and alternative media have started to take advantage of these inexpensive technologies to build networks of news gathering that shadow the large-scale operations of media goliaths such as CNN, BBC, and others." [6] Alternative news media in the

---

[3] Sean McLaughlin "The use of the Internet for political action by non-state dissident actors in the Middle East" First Monday, November 3, 2003. http://www.firstmonday.org/issues/issue8_11/

[4] For a comprehensive study of technological infrastructure in Africa, for instance, cf. Bandwidth Task Force Secretariat, *More Bandwidth at Lower Cost: an investigation for the Partnership for Higher Education in Africa*. Dar es Salaam: University of Dar es Salaam, October 2003.

[5] Reference: http://lanic.utexas.edu/info/newsroom/elections/
[6] Rafal Rohozinski, "Bullets to Bytes: Reflections of ICTs and 'Local' Conflict" in Robert Latham, ed., *Bombs and Bandwidth: the Emerging Relationship between Information Technology and Security*. New York and London: The New Press, 2003, p.306.

Palestinian territory on the West Bank regularly mount video footage of incursion events on the Web, to counter what they perceive as a pro-Israeli bias in coverage of the region in the major news media.

On the basis of this analysis of producers it is clear that on the Political Web persistence of content will be rare and patterns of change difficult to predict. In addition, important Political Web content can be dynamic and hence technologically complex to archive and preserve. It is also clear that Political Web sites are growing increasingly similar in functionality to on-line newspapers and media sites, as evidenced by the use of listservs to acquire newsletter subscribers by MIRA and the use of notification services by BurmaNet. These developments present significant technical and curatorial challenges for archiving the Political Web.

2. *User Behaviors and Needs*

The PCWA investigation was based on the premise that Web-based political communications serve as primary source materials for the study of political groups and events, much as print communications have served the same purposes for some time. Hence, archiving of those materials should support the "long-term availability for specific and limited educational and research uses" of those materials to serve two broad communities of interest:

1. *Scholars, researchers and teachers in the Humanities and Social Sciences.* These include scholars in a wide range of humanities and social science disciplines; Most are affiliated with colleges, universities and/or research centers.

2. *Researchers and analysts in the international development, policy, diplomatic, and journalism communities.* These include individuals engaged in research for the purpose of shaping and developing public policy, in some cases affiliated with government agencies, such as the State Department, the U.S. Congress; with international non-governmental organizations, and policy institutes.

To obtain in-depth knowledge of the interests and behaviors of the potential users of archived Political Web materials the investigators did two things: developed and implemented a survey instrument to gather information from a broad sampling of researchers; and convened and interviewed individual researchers from the academic and public policy communities about their use of Political Web materials as primary sources for research. See attachments 1 and 2 on *User Survey* and *Studies of Individual Users*

The research communities surveyed displayed distinct types of behaviors in their use of retrospective political Web materials, including project-based research, ongoing study, and occasional reference to or citation of Web sources. The importance of Web site persistence to the last, reference and citation, suggested that Web archiving should serve two broad purposes: preserving historical evidence for future research and providing persistent sourcing for current research.

The survey and interviews revealed a pronouncedly high interest among most researchers in the informational content of political Web sites, as well as in the discursive or ideological content or the artifactual qualities of sites that reflect political and social culture. Respondents to the survey and researchers interviewed indicated interest primarily in the textual content of sites. The study also suggested that adoption of on-line news sources by researchers has been rapid and that, while the evidentiary characteristics of paper newspapers are still significant in documenting original "instances" of news reports, the advantages of on-line news sources make them superior in some respects to paper sources for current research. Others studied indicated that newsgroups are rapidly eroding the centrality of the traditional news media, i.e., newspapers and broadcast channels, as the primary sources of news information from some regions.

Respondents also confirmed the curatorial team's belief that for many regions archiving of Political Web materials would be more valuable if done in conjunction with archiving Web materials produced by governments.

The researchers in the humanities, social science, and policy research communities studied displayed a range of behaviors in their use of retrospective Political Web materials. These behaviors included locating, gathering, archiving, analyzing, quoting, and citing data for the production of commentary, other knowledge products and, in some cases, policy-related reports. Users need to monitor sites over time, methodically gather content that may change from day to day, and detect and analyze changes that reveal evolution in agendas, shifts in message, and other changes in behaviors of the producing political groups. The authors of new knowledge products also require that documents and content cited as evidence and source materials persist and be viewable by readers as supporting evidence in the form in which they were originally viewed. The archiving of Political Web materials must accommodate all of those activities.

It is also significant that all three studies, though focusing on specific regions, required the use of materials that crossed national boundaries, in addition to those produced within the subject regions. Even where the focus of research is a single nation, materials generated elsewhere are often of great relevance. Hence harvesting of materials confined to a single country domain would not be effective.

Interrogation of these user communities during the PCWA investigation revealed a surprisingly high level of acceptance by scholars and policy researchers of Web-based materials, particularly on-line news, discussion list postings, and government sites, as source material. This phenomenon is also evidenced by the frequency of citation of Web-based communications in policy journals and in the literature of international studies.

Most researchers evinced a need to archive the fugitive materials for later presentation as supporting evidence, and had developed a variety of strategies of their own to compensate for the lack of a comprehensive archive with adequately sourced content. The amount of metadata about the sites gathered by these researchers, however, was minimal, consisting normally of URL, date and time accessed, and URL. This suggests that the structure of the sites, links between pages, the circumstances of their fabrication, and other artifactual characteristics are not highly valued by many researchers beyond their importance in preserving the integrity of the texts and the ability to associate those texts with their original source.

PCWA activities then should support two basic user needs:

1. *Historical analysis – The ability to track and compare instances of sites over time in order to chronicle significant changes in the activities, strategies and views of the producer groups, and to retrieve information and documents that those sites provided at a particular time.*

2. *Citation - The ability to use and re-present reliably sourced digital content "after the fact," i.e., from persistent archival repositories, to support analytical studies and discourse on political events and trends.*

While the number of studies that involve historical analysis is increasing, the citation of Political Web materials as "sourced content" is far more prevalent, and hence is an activity with a demonstrable and immediate need for support.

3. Existing Approaches to Archiving Traditional and Web-based Political Materials

The programs for archiving political materials in traditional materials provide some cost and organizational models potentially useful for Political Web archiving. The programs for archiving Web materials present opportunities for collaborative synergies with the PCWA effort.

How Political Materials in Traditional Formats are Gathered and Disseminated

Political communications issued in traditional formats, printed on paper or broadcast via analog radio and television signals, have long been collected and archived for purposes of scholarly and public policy

research. Much of this collecting has been subsidized by the federal governments and major universities of developed nations like the United States, Great Britain, France, and Germany.

Library of Congress OVOP: The most extensive and systematic collecting program for political materials is maintained by the Library of Congress. Under the Library's ongoing foreign acquisitions program and special projects like the PL-480 program and the Hispanic Acquisitions Project, the Library's Overseas Operations (OVOP) offices have collected traditional materials for the Library's own collections and for the collections of major U.S. universities. The Library's apparatus includes field offices in Latin America (Rio de Janeiro), the Middle East (Cairo), Sub-Saharan Africa (Nairobi), Southeast Asia (Jakarta), and on the Indian subcontinent (Islamabad and New Delhi). Through this apparatus the Library acquires and often microfilms newspapers, journals, and ephemeral materials like posters, pamphlets, and handbills with political content from the major regions of the world. The Library's staff obtains some materials directly from publishers, often through standing or blanket purchase orders, and others indirectly through dealers and agents who work from desiderata lists and specifications compiled by Library selecting officers and area specialists.

The Library makes these materials available for sale in the original or in microform copy to libraries in the United States and many libraries have standing arrangements to regularly acquire all of the materials collected by the Library in one or more regions. Some of the Library's costs are recovered through the sale of materials, but the major costs of the program are supported by federal appropriated funds. A number of other national libraries like the British Library and the major German research universities also operate federally-funded programs similar to the Library's for their own countries.

National Legal Deposit Programs: Copyright or legal deposit provides another mechanism whereby federal governments and their respective national libraries gather published and unpublished political materials. This mechanism covers domestically produced materials of all kinds, and relies on domestic producers to voluntarily submit their content, as one of the requirements for obtaining grants of certain legal protections of their exclusive rights to disseminate those materials. As a by-product of this activity the deposited materials become available for research use. While submission is voluntary in most instances libraries like the Library of Congress and the Bibliotheque Nationale de France accumulate significant portions of their domestic holdings through the legal deposit programs. Not all countries have active programs to administer and enforce legal deposit requirements, however. Because legal deposit is primarily intended to promote commercial publishing activity, however, political materials represent a relatively small portion of the materials obtained in this manner.

Center for Research Libraries: The Library of Congress and many major academic libraries at the large North American research universities also acquire foreign political materials through the Center for Research Libraries Area Microform Projects (AMPs). These programs preserve, largely through microfilm capture, political materials such as newspapers, governmental and private archives, and various kinds of journals and ephemeral materials from the major developing regions of the world. There are six AMPs, each devoted to a single region: Africa, the Middle East, Latin America, Eastern Europe, and South and Southeast Asia.

Each program is governed and funded by its members. Materials are selected for acquisition or preservation by program members, who are representatives of universities and research libraries that invest in the program annually through a membership fee. These representatives tend to be area studies specialists, bibliographers, and faculty. Content to be preserved is usually identified and preserved on a project by project basis. Materials acquired or preserved under the AMPs are then available for use by scholars at member libraries through interlibrary loan.

Another Center program devoted to preserving foreign political reports is the Foreign Newspaper Microfilm Project (FNMP). Under this program, critical foreign-language newspapers from developed and emerging parts of the world are microfilmed and archived for academic use. These activities involve agreements with the publishers, established media organizations, who provide copy for filming and permission to film, in return for the Center's providing them archived copy (microfilm).

Many major research universities like the University of Florida, Harvard, Princeton, Chicago, Cornell, University of Texas at Austin, and others operate individual or inter-institutional collecting programs as well. These normally focus on certain parts of the world on which the university has especially strong area and language expertise. These programs involve acquisition of newspapers, pamphlets, posters, and other ephemeral documents on specific topics and events on an ad hoc, project basis. They also involve ongoing acquisition of such kinds of materials through purchase and exchange agreements. Such programs are usually region-based, and rely on arrangements with publishers, book dealers, other universities, and national libraries active in the subject regions. They are normally funded through grants (special projects or endowment) or through library acquisition funds. Materials acquired are normally made available on-site at the universities at which these programs are based, but can often be purchase or borrowed in microform by partner institutions and others.

Commercial Re-Publishing Activities: Several commercial publishers also aggregate and preserve political materials for the scholarly research market. UMI-ProQuest, based in the United States, and IDC, based in Leiden in the Netherlands, are two of the largest such publishers. These firms acquire the rights to reformat and distribute copies of newspapers, journals, and government documents published in the U.S., Europe and the less developed parts of the world, and for archives and collections held by large institutions like the British Library and the Library of Congress.

In the U.S. the Foreign Broadcast Information Service (FBIS), a federal agency, has long recorded and disseminated radio and television news broadcasts and political reportage for the U.S. government, public policy, and academic research communities. The FBIS Daily Reports consist of translated broadcasts, news agency transmissions, newspapers, periodicals, technical reports, and government statements from nations around the globe. These media sources are monitored in their original language, translated into English, and disseminated on microfiche through NewsBank, a for-profit publisher specializing in news content. FBIS reports are also available online through the *World News Connection®* (WNC), a subscription-based product of the for-profit Thomson Corporation. The FBIS Reports are valuable resources for the study of foreign affairs, business, law, sociology, political science, and trade in all regions of the world. The market for these reports is the large research universities, policy institutes, and government agencies.[7]

The maintenance of the political content in these products by the commercial publishers, however, is driven by near-term market demand, and so is not dependable for the long-term preservation of that content.

How Political Materials in Electronic Formats are Gathered and Disseminated

Several initiatives exist that locate and harvest political materials from the Web and make them available for research purposes. These efforts tend to be either comprehensive, covering broad domains with very general selection criteria (the Internet Archive, PANDORA), or are specialized, focusing on materials on particular subjects areas (Wellcome) or events (George Mason University's September 11 archive). While none of these efforts adequately archive the entire spectrum of political Web communications, one or more of them might support the work of a comprehensive Political Web archive.

"Comprehensive" Web Archiving

The two most inclusive archives of the Web are maintained by the Internet Archives and Google. The Internet Archive, a not-for-profit corporation, periodically archives and makes available through its Web portal, the *Wayback Machine (*http://www.archive.org/), "snapshots" of the World Wide Web captured by the for-profit firm Alexa, with which the IA is affiliated through its founder Brewster Kahle. Alexa makes

---

[7] The BBC also monitors and extracts reports from political and media Web sites in the Middle East, South Asia, and other conflict zones, and distributes extracts in print and electronic form.

Web sites it has harvested and cached for its own purposes available to the Internet Archives. The Internet Archives in turn makes the archive available with limited functionality (searchable by text key words and by URL) to the general public gratis. This broad snapshot content, however, is not comprehensive even within a single domain. Since its content is gathered by Alexa for specific analytical purposes the Internet Archive does not consistently preserve Web site content in an archival manner. (An analysis of the Alexa harvests is provided in the curatorial and technical team reports.)

To date, the Internet Archive's archiving activity is largely sustained through philanthropic support from its founder, and is heavily dependent on the Alexa crawl activities for its content. More recently, however, IA has begun to provide specialized crawling and archiving services to clients like the Library of Congress on a contract basis for a fee. In these projects the Internet Archive has been able to achieve a higher quality of capture with its own focused crawls. The Internet Archive is also a technology partner to a Web archiving consortium formed recently by several national libraries, to explore electronic copyright deposit and archiving of Web content from the respective nations, and is developing an open source crawler for this project, called Heritrix[8]

Google, a for-profit Web search service, periodically caches the Web sites indexed by its search engine. Google takes a "snapshot" of each page examined as it crawls the Web and caches these as a back-up in case the original page is unavailable. The cached content is used by Google to judge whether a page is a relevant match for a query. Like the Internet Archive, Google's content is also not comprehensive and pages are removed or are not crawled when owners object. On the other hand the searching provided by Google for its own cached content has higher functionality than that of the Internet Archives.

Google is funded through advertising revenues, through licensing its search engine for use on Web sites, and through payments received for "privileging" some commercial sites and Web content artificially in its search results. Google is currently a privately held corporation, owned by its founders and a limited number of investors. Again, "preservation" of content by commercial organizations is market-driven rather than responsive to the research community and so is often short-term. Site content in the Google cache is preserved only for a short period, in some cases only a few days, and is constantly replaced by new instances of the site.

---

[8] Reference: http://www.crawler.archive.org

National Web Archiving Efforts

Within the past five years several national governments have begun programs to comprehensively harvest and archive Web content produced in their respective countries. These programs are generally undertaken by the national libraries, and stem from those libraries' mission to document the national heritage and to serve as sources of information for the nations' populace. In some instances they arise from the traditional role those libraries' play as national legal depositories.

The national Web archiving efforts represent significant investments by the national governments in digital preservation. The Swedish Royal Library endeavors to capture and archive Web content produced in Sweden, as designated by the .se domain, under its Kulturarw3 project. Australia, through its national library, initiated PANDORA, which seeks to gather important Web sites hosted in Australia.[9] (The technical characteristics of these harvesting and archiving efforts are discussed in the technical team report.)

The National Library of Sweden's Kulturarw3 program (http://www.kb.se/kw3/ENG/Statistics.htm) endeavors to harvest and archive all surface Web materials that are produced in or pertain to Sweden. The effort targets all Web sites with the domain .se and other Swedish Web sites among such top domain names as: **.or**g, **.net** and **.nu.** Kulturarw3 performs crawls of the Web which each last from one to eight months at intervals of one month.

The National Library of Australia harvests selectively, focusing on defined categories of Web sites, including government publications, university publications, conference reports and materials of current political interest.   Because the NLA harvests a predetermined, circumscribed list they can negotiate with every publisher for ingest and re-presentation; evaluate every site captured for its usefulness; and catalogue everything they harvest. (Every title is given a full MARC record in the NLA OPAC and the National Bibliographic Database. The selective model also allows them to check every title for completeness. (Roughly 40% of the titles need some sort of active intervention to make them functional.)

Both the Australian and Swedish efforts have generated a rich legacy of reports and statistics; examples of business and logical models; positions on general archival practices; and their approaches to specific issues of ingest, management, administration, preservation and access that can be applied to archiving Web-based political communications.

The Library of Congress takes a different approach to comprehensive Web archiving. The Library's Copyright Office is exploring the possibility of compulsory electronic copyright deposit as a means of collecting Web materials. The Library has also recently begun to solicit partnerships with parties in the commercial and non-profit private sector to develop strategies and methodologies for preserving the nation's digital heritage materials through its federally funded NDIIPP program.   Unlike the Australian and other national programs, the Library of Congress expects to rely heavily on the investment of the higher education and private sectors to support digital archiving. [10]

---

[9] More recently the PANDORA program has been scaled back to harvest selectively rather than comprehensively. The report, *Balanced Scorecard Initiative 49 Collecting Australian Online Publications,* recommended that the National Library should prioritise its collecting of online publications to focus on six categories:
- Commonwealth government publications
- Publications of tertiary education institutions
- Conference proceedings
- E-journals
- Items referred by indexing and abstracting agencies
- Sites in nominated subject areas on a rolling three year basis and sites documenting key issues of current social or political interest.

None of these categories is to be collected comprehensively and each will require selection guidelines to be developed in order to define clearly what the Library will collect.

[10] Recently, led by the Bibliotheque Nationale de France, a consortium of national libraries, including the Library of Congress, was formed to collectively explore and implement ways in which to archive Web content.

Some Local and Specialized Web Archiving Efforts

A number of Web archiving initiatives have emerged or are beginning to emerge in the United States, United Kingdom, and Europe, driven by communities of interested scholars, international development organizations, and librarians and archivists. The Wellcome Trust in England, for instance, has begun an effort to extend its collecting activities in the history and practice of medicine into the digital environment by establishing a pilot medical Web archiving project in collaboration with JISC, the British Library and the National Library of Wales. The project, which is currently in a two-year pilot, will evaluate the PANDAS software, draw up an ITT for the Web archiving infrastructure (on the model where resources and services are hosted centrally, and each partner institution engaging local Web archivists to identify relevant resources, negotiate permission to archive, and archive the sites. The Wellcome will focus on medical resources, NLW will focus on Welsh sites, JISC will concentrate on resources produced for the higher education community.

Other examples of topical Web archives are the Heidelberg University Chinaresource.org effort on Chinese Web materials (http://www.sino.uni-heidelberg.de/dachs/intro.htm) and the *Occasio Digital Social History Archive* (http://www.iisg.nl/occasio/) is an on-line archive of newsgroup messages on social, political and ecological issues from the Association for Progressive Communications (APC), an international partnership of communication networks. The messages are sent to the archive as they are generated. The archive is developed by the International Institute of Social History in the Netherlands. The Archipol Project (http://www.archipol.nl/english/project/projectplan.html) also involves the archiving of Web sites produced by political parties in the Netherlands.

Such specialized archives are characteristically supported by grant or philanthropic funds, or by the parent library or archives institutions as extensions of their traditional archiving and information-sharing missions. They draw upon resident subject-matter expertise provided by curators, subject specialists, and archivists.

Personal Web Archiving

Many of the researchers surveyed during the PCWA investigation expressed grave concern about the loss of Web content that served as evidence or supporting material for their research. Many took to "archiving" portions of the sites by saving text and image content using available software, such as Microsoft Word, HTTrack, Internet Explorer, and EndNote. The purposes served ranged from presenting the sites as evidence in publications and on-line works, to aggregating the content to personal or shared data bases. These softwares capture the characteristics of the digital content to varying degrees. Some preserve the look and feel of the sites; others retain only the text or image content.

In addition some researchers have begun to use repository-building software, like D-Space, produced by MIT, to create local or "institutional" repositories of their own research materials. However, the materials archived in D-Space tend to be largely secondary sources, such as conference papers, preprints of articles, or self-produced materials.

A Note about Subject Portals

Several large research universities have developed Web portals devoted to providing information for research on various regions of the world. Stanford University Libraries maintains the Africa South of the Sahara portal (http://www-sul.stanford.edu/depts/ssrg/africa/guide.html). The University of Texas, as part of its *Latin American Network Information Center (LANIC* http://lanic.utexas.edu/), maintains the *Electoral Observatory* portal, which links to political party and other sites relating to political elections in various parts of the Latin American world (http://lanic.utexas.edu/info/newsroom/elections/). The Library of Congress operates its global *Portals to the World (*http://www.loc.gov/rr/international/portals.html); and the United Nations Environment Program maintains the *United Nations Environment Network* (http://www.unep.net/), a global portal to authoritative environmental information organized around various

themes and regions *These portals do not archive Web content but rather provide subject access to that content which is currently "live" or maintained a an active site.* The value of the portals is in the aggregation of access to authoritative information on a region or subject and in the filtering out of unreliable Web content and sources.

The portals are normally available to the general public gratis without restriction, the cost of maintaining such services being borne by the organization as a mission-driven activity (United Nations) or by the university or university library (Stanford, LANIC) as a resource for local student and faculty research. A secondary return on the parent institution's investment is the visibility they provide for the organization's academic programs and resident expertise. Increasingly, universities are seeking outside funding for portal activities, either through grants or through offering derivative fee-based services related to the portals.

A variation on the free subject or region portals is CIAO, Columbia International Affairs Online. CIAO is a subscription-based Web resource that provides access to documents, articles, and published papers, many of them produced online by policy institutes like the Brooking Institution and others. To ensure persistence of some Web content re-aggregated by CIAO the service mirrors the content on its own servers rather than linking to it live at its original Web address.

4. Curatorial Regimes and Issues

Selection and Annotation of Political Web Materials

In general a Political Communications Web Archive (PCWA) would collect content disseminated via the World Wide Web by non-governmental political organizations and groups, by governments, and by alternative media organizations. Within this scope certain principles stemming from the intended uses of the archives should also govern selection. The archive should be politically neutral and inclusive, for instance, and should include political communications regardless of ideology, orientation, or level of controversy associated with a producer group or event.

Eligible for inclusion in the Archive are 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.[11] "Sites" would also include "Subsites," sets of Web pages within a site produced for or by a different entity than the publisher of the parent site, as well as "Supersites," typically a single Web site that extends over multiple network locations even though it is intended to be experienced as a single "place" by the user. These pages may include HTML files and all embedded or locally linked documents and files, including text, image, sound, and moving image files. Political communications often interweave symbols, pictures of historic individuals, colors, sounds, photographs, and other images along with text to convey their ideological message. Proper archival capture, then, would preserve all of the digital components for the homepage and all levels below it that share the same root URL.

Ideally, captured content should retain the "look and feel" of the original Web site. Curatorial team members, who included archivists, curators, and library area specialists, expressed a strong desire for this full capture.

Another critical functional requirement for the archiving activity is the need to preserve with the digital object the "authenticating" information that guarantees that the archival object presented to the end user corresponds in all important respects to the original instance of the site as captured. The importance of preserving the evidence of authenticity supports users who must reliably cite and reference Web materials

---

[11] "Web Characterization Terminology & Definition Sheet," W3c, http://www.w3.org/1999/05/WCA-terms/ This document also contains a useful discussion of the complexities involved in defining entities such as "Web site."

that have since vanished. This requirement conforms to the UNESCO Charter on the Preservation of the Digital Heritage.[12]

For the time being, capture of deep Web materials, such as interactive databases and password protected content, is not within scope of the archiving effort. Harvesting of such materials would involve a level of technical application and interaction with the producers/hosts requiring an investment of resources that could not, in the opinion of the project team, be justified on a cost-benefit basis. While such content may fall within the scope of this project, it presents formidable technical and cost obstacles.

In particular, the harvesting model proposed for Political Web archiving is based on a data harvesting approach, whereby a crawler visits sites "unannounced" and "pulls" the appropriate content into the Archive. Conversely, the most successful attempts to date to archive deep Web materials are based on a "push" model, where content providers work out arrangements before the fact with archive personnel, and then upload or deposit their deep Web materials into the archive. This level of cooperation from the organizations and individuals producing the political Web content, some of which is illegal in its country of origin, would be unlikely to occur.[13]

Selection categories are defined to allow for broad capture of not only materials produced by political organizations and activists, but also the political dialogues that take place within the political sphere. Specifically, this argues for inclusion in the archive of government sites, which disseminate policy, and alternative political media. Capturing the full extent of the political environment increases the archive's value as a research tool for comparative and historical purposes.

On the other hand, broad sweeps of entire domains, national or sector-specific (such as the dot-uk or dot-gov domain), would not be effective.   In many countries the gov domain is huge and much of its content is not pertinent to the PCWA,. Sites for ministries of economy, finance, treasury, and so on, often contain large statistical data sets, economic planning documentation, administration of services, and other materials unrelated to political activities. Under the alternative media rubric capture of mainstream or commercially produced newspapers would similarly be impractical. Many news organizations maintain, often by outsourcing, their own online archives of back issues, which are designed to provide a revenue stream. Archiving these would introduce complications with regard to copyright and impose constraints upon prospective business models of the Political Web archive.

As a "historical record" of political communications, however, a Political Web archive should be as inclusive as possible. Unlike the selection policy statements used by other types of Web-based collections–for example, by a subject directory or portal site -- where much emphasis is placed on quality and stability of the target site, the Political Communications Web Archive aims to capture as much of the political discourse as possible within its scope, ranging across the political and ideological spectrum and often disregarding such typical "quality filters" as the authority and reputation of the producers; accuracy of content; currency and frequency of updates; and aesthetics and design.

Selection Constraints: Copyright, Notification, and the Dark Archive

Legal restrictions on the use of intellectual property present obstacles to much archiving of Web content. Copyright applies to original works from the moment they are fixed in a tangible medium. For the PCWA project, that medium is the World Wide Web and copyright covers virtually everything placed there.

---

[12] Reference:
http://portal.unesco.org/ci/ev.php?URL_ID=13366&URL_DO=DO_TOPIC&URL_SECTION=201&reload=1070319074

[13] As one example of the difficulties and increased cost involved, harvesting deep Web materials without the direct participation of content providers would require Archive staff to manually fill out interactive forms in order to "generate" the content. This would need to be done on a case-by-case or site-by-site basis and would be extremely labor-intensive. In addition, some deep Web materials require costly software migration or emulation. Investigations into new techniques and approaches for harvesting and preserving the deep Web will be ongoing as part of the project, and it is hoped that an extension of the scope of the PCWA in the future would encompass at least some deep Web materials.

As stated in the feasibility study undertaken for the JISC and Wellcome Trust, *Collecting and preserving the World Wide Web*, ". . . it should be recognized that a significant element of the additional costs of the selective approach is occurred in rights clearance."[14] Characteristics of a political communications Web archive increase the difficulty of obtaining permissions. By definition the PCWA will not be a finite collection limited by an event or domain. The size of the collection over time is difficult to estimate. As an indicator, we calculated that 1,041 of the 14,400 unique links or URLs on LANIC conform to the project's definition of political Web sites.   Political communications content providers, moreover, deal largely with immediate issues, and long-term preservation of their material is not necessarily a priority for them. Sites that are event-driven tend to disappear within a year, possibly two, of that event, leaving a small window of opportunity to locate authors and follow through on obtaining permissions. If rights clearance were required, it would severely limited manageability of the project by restricting feasibility of capture and accelerating costs.

We obtained the advice of Georgia Harper, legal counsel for the University of Texas System and a leading specialist in copyright and intellectual property rights. Based on the analysis presented by Harper, we believe that the PCWA can claim exemption from copyright restrictions. We believe that this claim can be substantially defended and justified on the basis of a combination of the Copyright Act's provisions for Fair Use and for Library Privileges given the economic model the archiving effort adopts.

To promote transparency and conform to Web etiquette, we do recommend that a notification of capture be automatically sent by e-mail to the site contact at the time of capture. As part of this automated process, a list of sites without contact information would be generated and a follow-up attempt made to contact site authors. The inability to locate a site contact will not, however, exclude a site from the archive. Notification will include information about the purpose of the PCWA and contact information. This form of notification will implicitly create an opt-out option by alerting the site manager to why and by whom the site has been crawled. This is particularly relevant when overriding robot.txt blocks.

We are recommending overrides of robots.txt blocks in the interest of building a comprehensive archive and based on the fair use argument.

Archiving political communications is not without ethical concerns, as pointed out in a presentation to the curatorial team by Peter Lor, CEO of the National Library of South Africa. Dr. Lor cautioned that consideration be given to the misuse of archived materials by oppressive regimes that could endanger the lives of the creators. Again, notification to site authors and an opt-out option are policies to help mitigate this possibility.

The opt-out option is not intended to remove a site from the Archive, but to place it for a period of time in a dark archive that is not publicly accessible. To determine the blackout period for the dark archive, a review of practices was conducted. The results suggest that 20 to 30 years is a standard used by the various international organizations in their archiving to protect their (mostly self-generated) materials, with 50- to 60-year restriction periods for materials whose disclosure might harm individuals outside of those organizations. In most cases, there is no restriction by policy of materials that were publicly available when created. A 50-year blackout is recommended for PCWA materials that an author or site owner has specifically objected to making available. The 50 years for the blackout period would date from the initial capture of a site. At the end of the period, sites would be released to the light archive.

Curatorial Regimes

A consensus was formed early in the project that library and area specialists would play a key role in the selection process by providing seed URLs based on a general archives collection policy statement and guidelines. Harvesting the identfied sites would follow two timing patterns. One group of ongoing sites such as those produced by certain political parties, NGOs, and activist groups, would be scheduled for

---

[14] Day, Michael (2003) "Collecting and preserving the World Wide Web" A feasibility study undertaken for the JISC and Wellcome Trust, pg. 29. http://library.wellcome.ac.uk/projects/archiving_reports.shtml

periodic automated crawls. Sites whose content is likely to change in a less predictable way, for example in in response to events and cycles like elections or *coups d'état*, the schedule of capture would have to be customized by the selector according to a number of external and intrinsic factors.

Decisions on optimum timing and frequency based on intrinsic charateristics of the sites would be informed by data on content changes and other variables generated by crawl analysis tools such as those under development as part of Project PRISM.[15] PRISM researchers maintain that intrinsic charateristics of Web sites may signal potential threats to the integrity and longevity of a Web resource, including technological obsolescence, security weaknesses and breaches, human-error in developing and maintaining Web pages and sites, benign neglect, power and technology failures, inadequate backup and secondary systems. How these factor influence timing of harvesting will also be determined by the nature and magnitude of the losses that are acceptable to the selector and the archive management. (One factor in this will be the cost of absolute certainty.) Documented changes in the number or size of pages, structure, or format of Web pages and sites of interest may indicate risk, depending on the context. Iterative crawls, ongoing monitoring, tools and techniques to detect and assess change, and increased familiarity with resources over time all form the development of risk categories and appropriate responsiveness. The information about these factors obtained through the crawls and other archving activities, in turn, can inform the further work of the selectors.

The two timing approaches are not necessarily mutually exclusive. For example, a selector might send a request to increase the frequency of capture for a set time period leading up to an important election for a political party site that is already on the list for the periodic crawl.

To achieve a systematic approach, we recommend a distributed model for selection that can draw on the technical and area studies expertise found at different institutions. Guidelines that detail the qualifications and responsibilities of such participants must be developed cooperatively. The Archive would also benefit from piggybacking on existing targeting and selection activities located where expertise and capability are already concentrated and supported, in order to develop and maintain subject/region-specific portals.

Unlike the "automatic" or domain-wide harvesting approaches taken by the Internet Archive and the Swedish Kulturarw project, a selective approach can create a greater likelihood of quality assured holdings in the Archive. However, some sites would inevitably be missed with manual selection and it is difficult to predict the future needs of researchers. We investigated the feasibility of using the Internet Archive's Wayback Machine to provide a source of gross capture for later culling or retrospective harvesting, thus serving in a supplemental role alongside the project-generated crawls. There are frequent problems with missing images and broken links, inability to access navigational buttons, links redirecting to current instead of archived versions of sites, problems with Javascript, and problems displaying multimedia content.

A distinction between the crawlers used by the IA should be noted, however. The WayBack Machine uses crawl data donated every two months to the IA by Alexa, a Web search company. The Alexa crawl is programmed to meet its own business priorities. IA can add URLs to the Alexa crawl, but they do not have input into how the crawl is specifically configured. IA uses Alexa for larger crawls, but does have its own crawler to do more focused crawls, such as for the Nigerian elections sites. (Appendix 3)

Depth, Breadth, and Frequency of Capture

From the curatorial standpoint, an archived site should preserve the "look and feel" of the original. The depth should encompass the complete Web site, i.e., all Web pages, including embedded image, motion, audio, and other files, having the same root URL as the homepage. The Technical Team recommends the capture of near files as well -- those files that are necessary to make a page display, but may reside on

---

[15] Kenney, Anne R., Nancy Y. McGovern, Peter Botticelli, Richard Entlich, Carl Lagoze, and Sandra Payette (2002) "Preservation Risk Management for Web Resources: Virtual Remote Control in Cornell's Project Prism." *D-Lib Magazine*, Vol. 8, No. 1, January http://www.dlib.org/dlib/january 02/kenney/01kenney.html and http://www.dlib.org/dlib/january 02/kenney/kenney-notes.html
Masanès, Julien (2002) "Towards Continuous Web Archiving: First Results and an Agenda for the Future." *D-Lib Magazine*, Vol. 8, No. 12, December http://www.dlib.org/dlib/december02/masanes/12masanes.html

another server. These include linked graphics, javascript, or style sheets, as well as downloadable objects such as sound files, zip files, or pdfs.

Regarding breadth, the capture of external links is not recommended. The PCWA will be based on a selective acquisition model, and in many cases external links should be reviewed in the process of the initial site evaluation. Linked sites that fit within the scope of the Archive can be added to the crawl list on their own merit at that time. The link to the external site will exist in the archived page; the researcher can then copy the URL for search of the site in the Archive or WayBack Machine. (Documentation on the PCWA interface would explain this option.) From a technical standpoint, capture of externally linked content/sites is problematic from the standpoint of crawl and retrieval technology.

As previously stated, frequency of capture will best be based on a two-tiered system for periodic fixed-schedule crawls and time-sensitive crawls. We suggest that the crawler selected for the project be a smart crawler, which is, for example capable of using a Last Modified Date time stamp to return the list of URLs that have been updated/inserted since the previous crawl.[16] This information would be used in conjunction with crawl analysis tools to develop a timing and frequency matrix. The values in the matrix might be refined on an ongoing basis based on empirical data recording the actual frequency of content updates on these sites as generated by a "crawl analysis tool." Such a tool would scan the URLs on the periodic crawl list at a fixed interval and report on content updates at the target URLs. This information would then be used to update the matrix.

We conclude that the frequency of capture must be indexed to both the typology and technical characteristics of the target sites or domain.

Generating Metadata

For robust search and retrieval of materials in the Archive by end users, 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.[17] A metadata system that integrates all three categories into an extensible, flexible package is the end goal.

The two descriptive metadata standards considered were simple Dublin Core (reference: http://dublincore.org/documents/dces/) and MODS (Metadata Object Description Schema) (reference: http://www.loc.gov/standards/mods/) (For comparison, reference: http://www.loc.gov/standards/mods/dcsimple-mods.html). Both are XML schemas and OAIS compliant. Dublin Core was developed as part of the Open Archive Initiative to provide a metadata standard for cross-domain information resource description and has growing global acceptance. MODS was developed by the Library of Congress to carry selected data from existing MARC 21 records as well as to enable the creation of original resource description records. MODS records were created for the MINERVA project's Election 2002 Web Archive.

We decided to opt for MODS for three reasons: the Archive would conform to the metadata standard set by the Library of Congress for Web archives; MODS is richer than Dublin Core and permits distinguishing among various roles in authorship; and it works well with METS (Metadata Encoding and Transmission Standard), which is an encoding format for descriptive, administrative, and structural metadata for objects in a digital library.

The utility of the site-generated metadata for annotation and indexing of archived Political Web sites is limited. Some tags convey little information about actual content. There are others with relevant metadata

---

[16] Oracle Technology Network (2002) "Ultra Search Crawler Extensibility API" http://otn.oracle.com/products/ultrasearch/index.html
[17] **NINCH** (2002) *The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials,* Appendix B. http://www.nyu.edu/its/humanities/ninchguide/appendices/metadata.html (accessed 03.10.2003)

that would be useful to retain. A site for a Colombian guerrilla group (http://www.eln-voces.com/) has the following meta tags:

> <meta name="description" content="Esta es la página Web del Ejército de Liberación Nacional de Colombia. Sómos una organización guerrillera que lucha por construir un nuevo país con justicia social y una real democracia.">

On the basis of this survey we concluded that most of the descriptive metadata will have to be input manually.

Creation of the abstracts will be the most time-consuming component of the cataloging or annotation process. In order to both standardize the abstracts and hold down costs, the abstract may be kept brief, i.e., two or three sentences, and would follow simple, formulaic, rules. The abstracts would ideally be in the language of the target site. In addition, taking a lesson from the WebArchivists.org, it may well be possible programmatically to generate a key and major component of the abstract from the site's primary sections as evidenced in the structure, in a navigational scheme on the Homepage, or both.

The programmatic extraction of a portion of the contents, as well as limiting the abstract to perhaps three formulaic sentences, might keep the costs of metadata generation manageable in the context of an Archive that will someday include tens of thousands of sites.

The browse function of the user interface will be constructed from the controlled vocabulary subject terms. This vocabulary should remain relatively fixed. Region-specific keywords can be supplemented as needed to respond to new political developments. A search function will operate on all descriptive metadata fields.[18]

Cost Considerations

We have recommended the identification and selection process be distributed to take advantage of library and area studies expertise across institutions. By this same reasoning, the input of descriptive metadata does not necessarily need to be tied directly to those involved in the selection process.

Use of region-specific portal sites in the selection process has been suggested to leverage the sizable investment already made in site evaluation and selection. Costs are evaluated for the portals *Africa South of the Sahara* and *LANIC*. Portal-based selection of URLs for the PCWA would be performed as part of their ongoing identification and evaluation process. This analysis considers only labor costs. It does not include one-time setup costs, hardware, software, programming, system maintenance, or overhead. Analysis of labor costs for portals serves as an indicator of the cost of site selection and monitoring only. Our analysis placed the labor cost per site of selection and maintenance, exclusive of equipment, technical at about $1.20. With a large archive or political web sites this selection cost would be significant. If combined with portal maintenance and development substantial economies could be achieved.

The WebArchivist.org was contracted by the MINERVA group to develop a system for MODS catalog coding, as well as training and supervising coders. For the Election 2002 collection, the average cost of MODS coding for a site was an estimated as $1.50.

It would be practical to set up input "hubs" at institutions that have area studies programs, such as those designated by the Department of Education through Title VI of the Higher Education Act (20 U.S.C. 1121 et seq.) as National Resource Centers. There one can recruit from the body of international students such programs tend to attract. The cost for inputting the descriptive metadata based on the full-time salary for a student assistant is $2.40 per site to complete a record or $2.88 if fringe had to be included. This cost

---

[18] Reference: http://lanic.utexas.edu/project/crl/datainput.html

rises to $7.62 per site if done by a professional librarian and to $9.89 per site if fringe had to be included as well.

There should also be guidelines and/or agreements laying out the attributes and expectations, or requirements for accreditation, of contributing selectors, catalogers, and institutions. Items to include in a subsequent phase of this investigation are determining the list of region-specific keywords and drafting guidelines.

General Curatorial Methodology

The organizational structure of the PCWA, whether as a separate entity or a unit within an ongoing operation, has not yet been determined. However, the proposed curatorial practices present some operational considerations. We recommend a curatorial process with the following basic steps:

- Area specialists select sites for inclusion in the PCWA.
- URLS are sent with timing recommendations to a centralized source and added to the crawl list.
- After each crawl, MODS records containing the programmatically populated metadata are sent to in putters.
- Using a Web interface, descriptive metadata fields are checked and completed by in putters.
- The completed record is added to a metadata repository database.
- Records can be retrieved for updating, such as adding selector-generated annotations.
- Once a record is set, it is moved to a production database that generates the METS.

There are several components to the process, and the ideal model would have a centralized management to ensure quality control, commitment to growing the Archive, monitoring of workflow, and ongoing report review and evaluation. Following the staffing model for PANDORA or MINERVA a project coordinator or manager could provide this oversight.

A more difficult question is how the participating institutions in site selection operate. There are basically three types of selection behaviors that the model should accommodate: 1) institutional commitment to active search for sites to provide the bulk of regional sites on a weekly or monthly basis; 2) periodic submission of sites by motivated individual(s) within an institution; and 3) sporadic submissions coming from the user base. Level 1 might involve designating a lead institution which brokers selection of materials pertaining to a particular region. The submission process can follow one of two paths. All submissions are made directly to the central management; or levels 2 and 3 submit their sites to the lead institution, which checks them against the holdings in the Archive, adds timing recommendations as needed, then sends a compiled list for crawl.

Stable lead institutions will be important to ensure that a concerted effort is being made to build a comprehensive Archive and that event-driven sites are systematically collected. This level of operation would necessitate designation of staff time to the Archive and should not be done solely on a "goodwill" basis. Stipends, payments, or other incentives to participate should be provided to the institutions and individuals and should be identified as part of the business plan. Lead institutions could also be considered for taking on the task of hiring and supervising the data in putters for one or all of the regions covered.

5.   Technical Strategies

The findings of the technical team investigations, summarized in this report, were informed by the parameters and desired archive characteristics specified in the "Investigation Wire Frame" document. They were also shaped by the findings of the curatorial team and the responses of that team to questions posed by the technical team. The technical team's work involved evaluating the technical methodologies and tools used by a number of relevant, extant independent Web archiving programs. It also involved purpose-specific internal testing and analysis of available crawlers and other tools for producing data and the results that those tools yielded. For purposes of these analyses, and to provide a proof-of-concept for methodologies prescribed by the team for the annotation of captured sites and automated generation of metadata, NYU developed a METS viewer, which very successfully reassembled, presented, and allowed viewing and manipulation of the archived sites (see Appendices 35, 36, and 37).

The team prepared detailed evaluations of the following approaches, which are summarized in Appendix 13:

- PANDORA (Appendix 14)

- Kulturarw³ (Appendix 14)

- Internet Archive (Appendix 13)

- MINERVA (Appendix 13)

- WARP (Appendix 15)

Based upon the requirements defined by the long-term resource management and curatorial team investigations, the technical team evaluated these methodologies in the following categories: crawling methods, data storage model, data formats, archiving formats, and metadata captured. The individual analyses are provided in detail in the appendices to this report.

Because of the fluidity and complexity of the World Wide Web itself coupled with the volatility of the technology used to capture, store and preserve Web sites culled from it, a robust yet flexible architecture married to a metadata system that accounts for structural, descriptive, technical and administrative information is the key to managing these complex digital objects in order to assure their authenticity, completeness, long-term preservation and access.

Most harvesting projects/repositories embarking upon this task invoke the OAIS reference model[19] and the Trusted Digital Repository model[20] as the twin bases upon which to construct a viable system for preserving access to Political Web materials. In general, a Political Web archiving effort must incorporate an OAIS-compliant, trusted repository, which is modular, scalable, and tightly bound by a flexible, extensible metadata system.[21]

Crawl Data for the Investigation

To obtain a test bed of political Web communications for the analyses the technical team relied on two central crawls of content from our partners and advisors, the Internet Archive, with supplementary and comparative data from crawls completed at NYU and Cornell (see the Harvester Case Studies and the Nigerian Election Crawl results in the attached appendices). The Internet Archive contracted to provide

---

[19] See the standard references e.g. the CCSDS Blue Book document *Reference Model for an Open Archival Information System*, 2002, http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf;
Preservation Metadata and the OAIS Information Model, A Metadata Framework to Support the Preservation of Digital Objects, OCLC and RLG, 2002. http://www.rlg.org/longterm/pm_framework.pdf

[20] See *Trusted Digital Repositories: Attributes and Responsibilities*, RLG and OCLC, 2002. http://www.rlg.org/longterm/pm_framework.pdf

snapshots culled from larger Web crawls undertaken by Alexa from a list of seed URLs provided by the investigation's curatorial team. The crawls focused on Web sites identified by curatorial team members and advisors, which covered a wide range of types and many regions of origin.

The 100MB aggregate .arc files we received from these snapshots contained a farrago of pages from many different dates within a crawl period of eight weeks of a particular domain. The purpose of this exercise was to introduce us to the .arc format and provide us with a test bed of materials from which to evaluate Alexa harvests as a possible source for pre-selected, focused archival content.

In addition, the Internet Archive also undertook focused crawls daily for a month-long period of 38 selected Web sites from the Nigerian Election of April 2003, selected by Karen Fung, using the IA's own crawler. The resultant .arc files were organized on a capture date basis, with one day's worth of all 38 sites bundled together in a single arc. As the seedlist grew, more than one 100 MB .arc was necessary to package one day's crawl. Contrasts in the variety of packaging along with the inherent differences in a packaged Alexa crawl vs a focused IA crawl made for an interesting analysis of the varied capabilities of the Internet Archive as a content broker.

Harvester Evaluation and Recommendations

An essential component of any Web archiving endeavor is the reliability and appropriateness of its harvester. The ideal harvester should create a safe archival copy, preserving or documenting key curatorial characteristics of the site, and facilitate the re-presentation and access to the service version by copying or representing the directory structure of the original Web site. It should package and contain the archive as hermetically as possible. To avoid linking out from the archived version to the live version of the site, the harvester should translate absolute links to relative links that reflect the storage path on the storage file system. It should weed out duplicate files and should switch off such functions as mail-to's, payment devices, external links (if desired), and forms. Finally, it should have the facility to perform both repeated sequential full snapshots and an initial snapshot followed by incremental, self de-duping harvests, and analyze capture to ascertain risk factors and create frequency algorithms.

The technical team had access to a number of harvesters and their results through its member organizations that contributed to its harvester evaluation. The team also gathered essential information about the practical application of harvesters, which are presented in the harvester case studies. Detailed results of the harvester evaluation, the set of harvester requirements, and recommendations for selecting a harvester for political Web archived appear in Appendices 16-19. The results of the harvester testing also populate many of the other appendices of this report.

Four harvesters analyzed in depth were:

1. The Alexa crawler

2. The Internet Archive's focused crawler

3. The NEDLIB harvester

4. HTTRACK/PANDAS

The Internet Archives open-source crawler, Heritrix, is a strong contender for use by the PCWA. It will always be necessary to seek, test, develop, adapt, and extend available crawlers as the nature and content of the Web and its enabling technology evolve.

Metadata

The curatorial team addressed the appropriate general requirements and production regimes for descriptive metadata, leaving the Technical Team to take on administrative and structural metadata. For this portion of the investigations, the technical team undertook the following activities:

- A feasibility study of automated harvesting from crawler logs (see Appendix 20). If metadata exists on Web pages, it can be harvested, but very often the metadata is not included, as further documented in Appendix 34, a comparative analysis of page data using the Nigerian elections sites.

- An evaluation the potential of METS for storing and delivering archived Web sites (see Appendices 21-24). This evaluation led to the development of a METS prototype that adapts page-turner functionality for the presentation of Web sites. The prototype produced very promising results for searching across and between multiple instances of selected sites over time.

- A mapping of the .arc file format to OAIS preservation metadata categories. As Appendices 25-27 illustrate, the .arc format incorporates a small proportion of the preservation metadata set.

- A robots.txt evaluation that considered the prevalence and potential impact of the use of robots.txt by test Web site administrators (see Appendix 28).

- An analysis of meta tags use on behalf of the curatorial team that is further discussed in the Curatorial report, and described in Appendix 29 of this report. This led to a spinoff evaluation of the nuances of meta tags described in Appendix 31.

- A review of Title metadata, described in Appendix 30 that raises serious concerns about the reliability of this metadata for automatic harvesting. These results are also referenced in the Curatorial section of this report.

- Metadata was also a focus of the evaluation of current methodologies (see Appendix 13).

- Ongoing monitoring of the work of the PREMIS working group on preservation metadata (http://www.oclc.org/research/projects/pmwg/) and its potential implications for Web archiving.

A General Conclusion Based on These Activities

The technical team's analyses determined that the overall approach of the Internet Archive was most closely aligned with the parameters for operation outlined for the Political Web project, being the most flexible, generally the least expensive approach, increasingly open source, and benefiting from ongoing, incremental, modular development that harnesses and initiates technological developments to enhance its capture, storage, and access approaches. The Internet Archive's recent beta release of a more comprehensive access interface goes a long way to eliminate some of the limitations that surfaced in the analysis of the test bed crawl results. The Internet Archive's Heritrix harvester is an open-source, java application that leverages the lessons learned in the development of Alexa and Mercator crawlers, and is more scalable than NEDLIB, which relies upon a database back end.

The .arc format in which IA saves Web content is a "lowest common denominator" format, a large zipped file containing other files. It contains three distinct sources of metadata: the file header for each file collected (containing info like IP, timestamp, mime type, and file size), HTTP headers, and then the file content itself which can be mined for further information. The accompanying .dat file effectively parses out useful metadata into a field-value list that can be easily processed by other applications.

Some archiving processes and activities will have to be automated to reduce the costs and complexity of the overall archiving endeavor. Such activities include content capture, production of descriptive and technical metadata, and annotation and cataloging of the captured digital objects. While automation of the content capture itself has reached an advanced state in the harvesting efforts studied, further work is needed on the development of tools for retrieving and generating metadata and annotation, in order to realize similar economies in those activities. Developments in this area like METS profiles and other applications, Internet Archive tool development, the Nordic Web Archive (NWA) toolkit, JHOVE, and others promise to yield benefits for Web archiving work.

Digital Preservation Approaches and Capture Considerations

From a technical perspective, one requirement for the archiving specified by the curatorial team presents particularly thorny challenges for digital preservation. That requirement is the need to preserve the "look and feel" of the captured Web sites. This would require that the Political Web archiving activity accommodate the full range of formats generated by sites' producers.

The file formats that predominate on the Political Web test bed sites, for the most part, present fewer preservation problems than other types of digital collections because they are primarily text-based formats, mainstream image formats, or other widely-used formats. (See Appendix 34 for detailed MIME results from a review of Web crawls for the project test bed sites.) However, some application-dependent formats and other types of formats that do not yet have defined preservation pathways do occur. And there will surely be new formats for which preservation approaches must be identified.

A policy of not limiting the acceptable formats, as other Web archiving projects surveyed have done, would have direct operational implications for the archive. If the political Web archive is to accommodate all file formats in use by the producers of Political Web content, it will have to take a format-specific approach. This would be a significant cost factor in the data management, storage, access, and perhaps other archiving activities outlined in the proposed model. And, since there are accepted approaches for some file formats that have proven preservation track records; some good management techniques for other formats that are harder to preserve; and no known approach for some new, complex, or extremely software-dependent formats, this requirement would introduce a great deal of uncertainty into the archiving cost model.

The technical team explored four different options for defining levels of preservation. Two of the four options would strike a balance between accommodating curatorial concerns and minimizing uncertainty.

1. Accept all file formats submitted then assign preservation level categories by formats to make explicit the extent to formats will continue to be available over time: e.g., "this digital archive will provide full level one preservation for all text-based formats for an unlimited time period, and level three bit preservation for x type of application format for the next five years with review at that point."

2. Accept all file formats submitted, and then convert selected formats for which no preservation approach exists or that are not widely-used to one of a limited set of preservation formats as determined by the digital archive. This is an inclusive approach that while ensuring persistence may entail loss of some functionality.

Whatever the approach chosen, it will be important to be explicit about the policy the digital archive will adopt towards file formats that do not yet have defined preservation solutions at the time of capture.

6. Sustainable Archiving – How Best to Organize, Govern, and Fund the Activities

Long-Term Management of Archived Resources

The task of the long-term resource management investigation was to determine how the archiving activities can best be self-sustained. This entails identifying:

1. the comprehensive set of activities required to preserve Political Web materials and make them available to the scholarly community
2. what resources, monetary and non-monetary, are required to support those activities on an ongoing basis; and
3. how to bring those resources to bear on the archiving effort.

The investigation analysis addresses resource requirements and economic sustainability, accountability to the user communities, transparency, and other appropriate criteria. The proposed model presents the configuration of activities believed most likely to support the ongoing archiving of Web-based political communications. As the technical team report noted, funding for preserving such materials in digital form to date has been largely episodic or sporadic, aside from a few national efforts in older developed countries like Denmark, Sweden, and Australia.

A fact that informed our conclusions and shaped the proposed PCWA model was that scholars are only beginning to accept Web materials as primary, citable historical evidence of evidence of political, social, and economic trends. The extent to which scholars will employ retrospective or archived site content -- as opposed to active sites -- in their work, and how they will employ that content can be surmised only on the basis of the experience of a relatively small number of "collectors," and on how analogous material in print

form is used. It is, moreover, uncertain how quickly printed materials will be replaced by digital ones as primary sources for research in the political sphere.

One might argue that retrospective political Web materials, like the collections of foreign-language printed ephemera and political publications that now reside in libraries, will probably be used infrequently. To be sustainable in a "low-demand market," the proposed PCWA model is designed to be aggressively opportunistic, capable of building on existing local or specialized, even commercial, archiving activities. Archiving activities on this model exploit capability and capacity wherever they both exist and are supported by stable, mission- or market-driven programs. Specialized "local" activities, such as selection by university, library, and government area studies specialists, archiving under copyright deposit programs at national libraries, and indexing and Web caching by commercial organizations, will be an important component of a viable archiving effort.

On the other hand the phrases "used infrequently" and "low-demand," which are useful in characterizing physical library collections, might not be meaningful when applied in the digital realm. The broad constituency and wide range of uses opened up by this type of resource might indeed translate at some point into additional sources of revenue. As conceived an archive of political Web materials would aggregate a great deal of material in one "location" for broad and global user access. This in turn expands the definition of the type of research that can be done using the archives, such as country risk analysis by multinational companies, financial and brokerage firms, and government agencies.

While the issue of "replacement" has budget implications in terms of being able to shift acquisition funds, from the researcher's point of view Web materials should probably be viewed as a supplement, rather than a replacement, for currently available research materials, just as JSTOR enabled many new uses for journal content that had been widely available in print for years. Rather, in terms of scope and scale, but also in terms of methodology and substance, it is more a matter of new kinds of research being enabled by this type of resource.

As one curatorial team member noted, "This is the essence of what has made Google so powerful and really brought the Web to bear as a first source of information for the general public. . . There are two good test cases of this in the commercial market today underway. Amazon is digitizing all books in and out of print- on the assumption that by making out of print books available and their texts easily searchable they will sell more books overall, even ones that are no longer in print, since people can fine tune what they are looking for."[22]

Essential Political Web Archiving Activities

Preservation of the important political materials on the Web will require a framework wherein a large set of activities can be undertaken on an ongoing basis, and that enables these activities to be both self-sustaining and responsive to the needs of the user community.

For purposes of modeling such a framework, a set of activities is listed below. These would enable the assembly, preservation and accessibility of a persistent and inclusive archive of Web-based political communications.   Some of the activities are generic, or broadly applicable to archiving all Web-delivered materials. Others match the specific needs of users and characteristics of political Web materials.

a. Selection / Curation

- o   Prospecting for archives content

---

[22] A second commercial pilot case is Netflix, an online DVD rental service which allows browsing ad searching millions of films by subject, main character, date and so forth. The added functionality has dramatically increased Netflix sales outside of the traditionally heavily used top fifty titles.

- Content selection / identification of defining characteristics (e.g., URL/domain, creator, topic/content, type) of target materials / "peer review"

- Determination of frequency, depth, scope of capture

- Certification/documentation of authenticity of initial content

- Indexing / metadata production / cataloging

b. Stewardship / Brokering

- Determination and monitoring of archives scope

- Creation / specification / adoption of criteria and standards for archives content

- Authorization of selectors and cooperating archives

- Asset management – rights, funds, resources

- Management / re-aggregation of archives content

c. Ingest / Harvest

- Pointing / programming of Web crawler

- Executing Web crawl

- Capture of content and metadata

- Notification of content producer re archiving

- Incorporation of cooperating archives' content

d. Administration / Data Management

- Development and procurement of enabling tools and technologies

- Quality assurance / auditing of archives content

e. Secure Data Storage / Repository

- Storage of master copy ("dark archives")

- Storage of service copy ("light archives")

- Storage of fail-safe copy (backup archives)

- Storage / maintenance of bits

f. Access / Interface

- Structuring presentation of content

o   Presenting content for viewing and searching

o   Authentication of users

How Best to Organize the Archiving Activities

The archiving effort can be organized or configured in three general ways:

1. *Distributed or "peer-to-peer,"* where content selection and management activities are undertaken by parties operating independently, using a variety of tools and standards. Archives based on a distributed model are created, for example, through the OAI Metadata Harvesting scheme and institutional D-Space archives. This model also includes peer-to-peer sharing of digital content as well, on the Napster model, as exemplified by the Herodotus system at MIT.[23]

2. *Federated,* where some important activities, such as selection and storage, are distributed and others, such as data administration and management, are centralized, creating and maintaining a common resource (e.g., metadata production for OCLC; production of metadata and content for Research Libraries Group *Cultural Materials* resource); MIT is also exploring formation of a federation around its D-Space software, where tools and standards are developed and disseminated centrally and are then supported by a community of users. Successful examples of the federated development of shared data resources abound in the natural resources world. One such example is the World Resource Institute's Global Forest Watch, whose local partners around the world supply data that is up linked to the database via satellite.

3. *Centralized,* where all important activities are performed by a single party or organization (Library of Congress *107th Congress* Web Archive or the *Africa South of the Sahara* portal).

Because of the obvious practical drawbacks and high risk associated with the last, only the distributed and federated models will be considered here for the Political Web archives. The entirely distributed or "peer-to-peer" model, moreover, does not address the issue of accreditation of selectors, which is essential to the integrity and reliability of an archive of record.

Many of the activities described in the model as they apply to developing and maintaining a common resource or archive, can be applied to creating local archives as well.

The recommended organization of the archiving activities is expressed in the proposed model using a "value chain" approach. Each activity is listed with its outputs, participants and their characteristics, accountability, and general requirements. These factors vary from one activity to the next. For instance, core activities like Selection and Management must be highly sensitive to, and hence controlled by the primary user community. Other functions, like Ingest, Data Administration, and Repository, are subordinated to core activities and might be outsourced to other entities that also serve other constituencies, such as government, commercial research organizations, and publishers.

Determining which activities are best centralized and which performed locally should be based on value. In general, activities should be performed centrally that benefit from the achievement of economies of scale that cooperative resource sharing can provide, or where the assets developed are significantly increased in value through aggregation. Such activities might include Selection, which requires a high level of specialized knowledge, such as uncommon language expertise or knowledge of a critical region. **In general, activities that are best supported locally should be performed locally.**

The Benefits and Drawbacks of Prospective Governance and Funding Systems

The funding mechanisms that support the critical archiving activities will affect the archiving effort's responsiveness to the user communities and will enable or impede strategic growth of the archives to a

---

[23] Cf. Timo Burkard. *Herodotus: A Peer-to-Peer Web Archival System*, Master's thesis submitted to the Massachusetts Institute of Technology, May 2002 paper, available at http://www.pdos.lcs.mit.edu/papers/chord:tburkard-meng.pdf)

greater or lesser degree. There are several funding systems that might be adopted to support one or more activities of Political Web archiving. Each system has distinct implications for accountability and sustainability of the activity.

> *Government / Entitlement* – National Web archiving efforts like PANDORA, Minerva, and Kulturarw3 are funded by the federal governments of Australia, the U.S., and Sweden respectively. These efforts benefit from the relative stability of appropriated funding, although federal funding levels are sensitive to a wide range of competing public needs, such as security, health, and education, and to constituencies far more populous and broad-based than the research communities of interest that would be served by the archiving of political Web materials. In creating a global archive or resource, moreover, funding from any single government is likely to favor the interests of one user community (e.g., the educators and scholars of that nation, or as with the Library of Congress its legislature) or the purposes of one regime, over those of others.[24] Hence the critical activities of a global effort cannot be overly reliant on funding from a single government.

> *Philanthropic* – The purpose of philanthropic funding is to catalyze new and promising initiatives, but not to maintain extant ones. Moreover, the interests of donor individuals and even organizations are likely to change over time. Hence this funding model is useful for the capacity-building stages of a program's lifecycle, but offers no guarantee of continuing support for the effort. Models: Internet Archives, Wellcome Trust.

> *Subscription or Access-based* – These systems are applied in both non-profit and profit-making contexts. Here the user community supports the archiving effort directly through payment for access to archives content on an annual or per-use basis. Users might also support the system by paying for inclusion of self-selected content/sites. Under such a system immediate demand and volume of use of materials tend to drive content and functionality and can distort long-term consistency and value. As JSTOR has shown, however, the subscription fee can be structured in such as way as to provide some assurance of stability and support of the archives' mission for the long-term, and insulating archiving activities like Selection and Management from short-term "spikes and dips" in user interest and demand. Models: JSTOR, RLIN, OCLC, e-journals.

> *Consortium* – Ongoing support of archiving is provided, and control exercised, by an organization representing the user community. Here organizations like libraries, universities, research centers, institutes, and agencies mediate the interests of the user communities, ensuring adherence to the collective interests of those communities while rationalizing and resolving individual, local, and other particularistic short-term demands. Models: BioOne, Center for Research Libraries Area Microform Projects.

> *Open* Access */ Producer-supported* – This model is a hybrid of subscription and consortium models. Here the producing university or organization subsidizes the initial production and preservation of content (usually scholarship) which is then made available gratis or for a nominal fee to an unrestricted audience. This model is being advocated for scholarly publishing by the Public Library of Science and other publishers of open access journals.

Of the aforementioned systems, the consortium model is the most likely to promote the optimal level of accountability to the primary user community, the greatest stability and likelihood of persistence over time, and the greatest potential for access to the wide range of competencies and capabilities needed to maintain broad-based, common archiving activities. An archiving consortium could be constructed in a number of ways, each representing the user community in its own manner. An archiving consortium of the world's major academic universities and Humanities and Social Sciences research institutions, for instance or of the major scholarly societies might serve all sectors of the higher education and policy research communities. The Center for Research Libraries will function as the locus for the Management activities or layer of the Political Communications Web Archiving effort.

---

[24] One might conclude this based on LC selection of topics for its Minerva Web archiving projects, such as September 11, Elections 2000, 107th Congress, and Iraq War, topics which were sensitive to broad national security and political interests.

The recent and continuing decline in allocation of university funds to acquisition of primary source materials for the humanities and social sciences, as opposed to STM journals, and materials not related to North American and Western European studies, is a factor that must be reflected in the model for political Web archiving. The Center's Area Microfilm Programs have provided a low cost model for cooperative collecting in this area in the analog realm by distributing selection activity to take advantage of expertise supported locally; having Ingest activities undertaken by commercial partners when appropriate; and having the financial support of the effort under the control of the chief stakeholders, i.e., the area studies departments of the participating libraries. The PW model posits a similar distribution of activities and resources.

Finally, in analyzing the costs and benefits of a prospective funding system due attention must be given to non-monetary, as well as monetary, incentives and bases for exchange. For instance, participation and services that support the archiving activity might be compensated by access to functionality and tools provided by the central archiving organization.

Economic Aspects of the Model

The development and support of an ongoing Political Web archiving effort would be undertaken on a cooperative basis, supported by the user community and its proxies.

Problem: The challenge is one of timing and resources. Universities and their libraries are already stressed by the current costs of supporting research in humanities, social science research. With the increasing emphasis on, and costs of, resources for core undergraduate curriculum and English-language studies in the areas of Science, Technology, and Medicine, the acquisition of foreign language materials has declined in most universities; as has the maintenance of foreign language and area expertise.[25]

Solution: The model proposed here has multiple potential revenue streams and forms of support. It depends upon many of the separate component activities of the archiving effort being supported by various stakeholders/parties, who are motivated by various kinds of incentives that serve their own local self interests and thereby earn their (local) support of the activities that generate the common goods of the central archiving endeavor. The forms of support may include, apart from funding, content, expertise, functionality, and other assets and resources that represent valuable bases of exchange.

In some cases the central archiving effort will build upon existing activities already supported at the local level, i.e., by universities, institutes, and even individual researchers. These activities include library collection development and selection and portal development. The Library of Congress *Portals to the World* project, for instance, is supported by the Library's subsidization of selection and annotation work by LC subject and area specialists and Federal Research Division specialists. This work and expertise identifies sites that might be periodically captured and preserved by the Political Web archiving effort. In such a scenario the Library would then be subsidizing selection for the Web archives.

Similarly Stanford University populates and maintains the Africa South of the Sahara portal, and spends over $75,000 per year in selection, maintenance, and Web support of that effort. The university's investment is justified by the benefits to the local community and by the exposure that the portal earns the university among the larger academic and international public policy community. These incentives justify the university's continued subsidization of high-level selection and monitoring of Web-based content in an important region. For a central Web archiving effort this work could provide two tangible benefits: authoritative selection of important content; and data about the persistence and functionality of that content, which could further inform and refine selection for the archives.

Other incentives for participation by selectors could be feedback about the persistence and behavior of specific Web objects or types of Web objects, in the form of data derived from the capture results of the

---

[25] A countervailing trend is the increased emphasis in the undergraduate curriculum on primary source research and inquiry-based learning; as well as the heightened focus in the public policy community on intelligence on foreign affairs.

central archiving activities. This feedback might provide valuable new information about the nature and rate of change in Web site content, building upon and refining for instance the general risk assessment information that Nancy McGovern and the Cornell partners developed based on the technical characteristics of the sites. Such feedback would then inform and enhance the capabilities and effectiveness of local selection, thus providing a local benefit.

Similarly the central Web archiving effort might provide more individualized services to individual researchers or their sponsoring organizations or publishers by capturing, archiving, and making available for continued presentation Web content which they cite in research products. The PCWA would then ensure the continued availability of the content, and its evidentiary integrity, adding value (and validity) to their product, in return for a fee. A pricing structure adopted for such a service might involve an initial capture fee and a smaller ongoing maintenance fee. Both fees would be keyed to the complexity of the digital object archived, and other factors such as if a licensing fee had to be paid to the producer of the site. (Secondary revenue streams could come from providing the same archiving services to the producer.) The incentive for support from the research community would be in the enhanced credibility of their publications.

A second potential funding strategy, but one that will "come on-line" slowly, is to draw from library and institute budget lines for activities, such as newspaper subscription, preservation, service, and microfilming, for which the archiving activities will provide a viable substitute and which they will perform more effectively. This would emphasize preservation of materials at that end of the Political Web spectrum that is heavy in news or information-dissemination content, rather than a proselytizing content. As indicated in the Production section, above, political Web sites share many behavioral and technical characteristics with the on-line newspapers issued by traditional commercial media organizations. The content of both kinds of sites, highly sensitive to political events and cycles, is likely to follow similar patterns of change, thus requiring comparable selection regimes. The user survey and researcher interviews indicated, moreover, a higher priority on use of this kind of site than those more heavily dedicated to advocacy and partisan activities.

A secondary revenue stream might also be derived here if the Web archiving activities provided services of value for the news-producing organizations themselves, such as certain preservation services that would ensure long-term availability of content useful to the producing organization, or distribution of retrospective content to secondary markets. In such a scenario support might then come from the producing organization, in the form of contract for service, and/or the end users in the form of pay for view or subscription. The archiving effort might then be of value as an archiving and distribution mechanism for the news producers.

This second funding strategy would have natural linkages to two of the Center's established area studies resource development programs: the Area Studies Microfilm Programs (AMPs) and the International Coalition on Newspapers. These two programs focus heavily on preserving news content from outside the United States. These linkages would no doubt yield efficiencies and savings.

Costs

Project participants at Cornell University developed a conceptual model for the digital preservation management workshop (http://www.library.cornell.edu/iris/dpworkshop/) as a starting point, and identified cost areas that needed more investigations within that model.

The technical infrastructure costs for collaborative Web archiving (e.g., cost, human resources needed to operate, server capacity required to run, storage considerations of output) will be influenced by the choice of crawler; the distribution of personnel across the collaborative enterprise (e.g., location, seniority); overhead factored based on participating members plus central unit, if appropriate.

Startup and ongoing costs are by definition more quantifiable. The proposed PCWA model provides general cost factors and principles, which are augmented by some specific costs provided in the curatorial team report and the technical and curatorial appendices. In general costs will include startup and capitalization costs; ongoing operating costs; and variable costs that are affected by volume of content, complexity of content selected, nature and amount of functionality provided, and volume of use. Additional data will be gathered in the next phase of the PCWA effort.

In general, the degree to which the activities can be automated will affect costs. Our analysis of the activities functional requirements and of curatorial practice in the analog domain suggests that as time passes an increasing number of activities will be automated. Combined with the fact that storage of the content, a major cost, will decrease in cost, the rise in costs of increasing amounts of content will be at least partially offset.

The methodology evaluation also considers program costs and the harvester evaluation includes cost implications. We acknowledge that even open source crawlers have associated human, equipment and other costs to incorporate.

To contain costs, the proposed model permits archiving of the Political Web to be implemented incrementally. This could take two routes. Archiving could be begun by initially limiting capture to textual content and relatively simple digital objects, which would reduce programming and other data management costs, storage costs, and would reduce uncertainty about long-term preservation. Second, archiving could be undertaken for a single or limited number of research areas, reducing selection, annotation, data management, and storage costs. Of the two choices the overwhelming favorite of the curatorial team curatorial team was the second. The next stages of the PCWA endeavor outlined in *Section 8* of this report reflect this feeling.

7. A Proposed Service Model for Political Communications Web Archiving

The illustration below provides a functional model for the Political Communications Web Archives (PCWA) as envisioned and specified in the preceding reports. The proposed PCWA model enumerates the individual activities or "layers" of a distributed ongoing effort to preserve important content from the Political Web.

Each of these activities is described in the text of this section of the report. The description includes four elements for each activity or "layer" of the model:

1. *Functional requirements*: the activities, processes, and outputs of the activity or "layer"

2. *Participants:* the general characteristics, skills, and capabilities of the individuals or organizations undertaking the activity.

3. *Cost factors and sensitivities:* the general types of costs and the factors that influence the cost level for the activity and incentives for investment by participating organizations and entities.

4. *Accountability and control:* the organizations or constituencies to which the entity performing the activity is accountable, and which exercises control of the inputs and outputs of the activity.

The activities in the PCWA model map to the main functions in the OAIS Functional Model. The OAIS functions are:
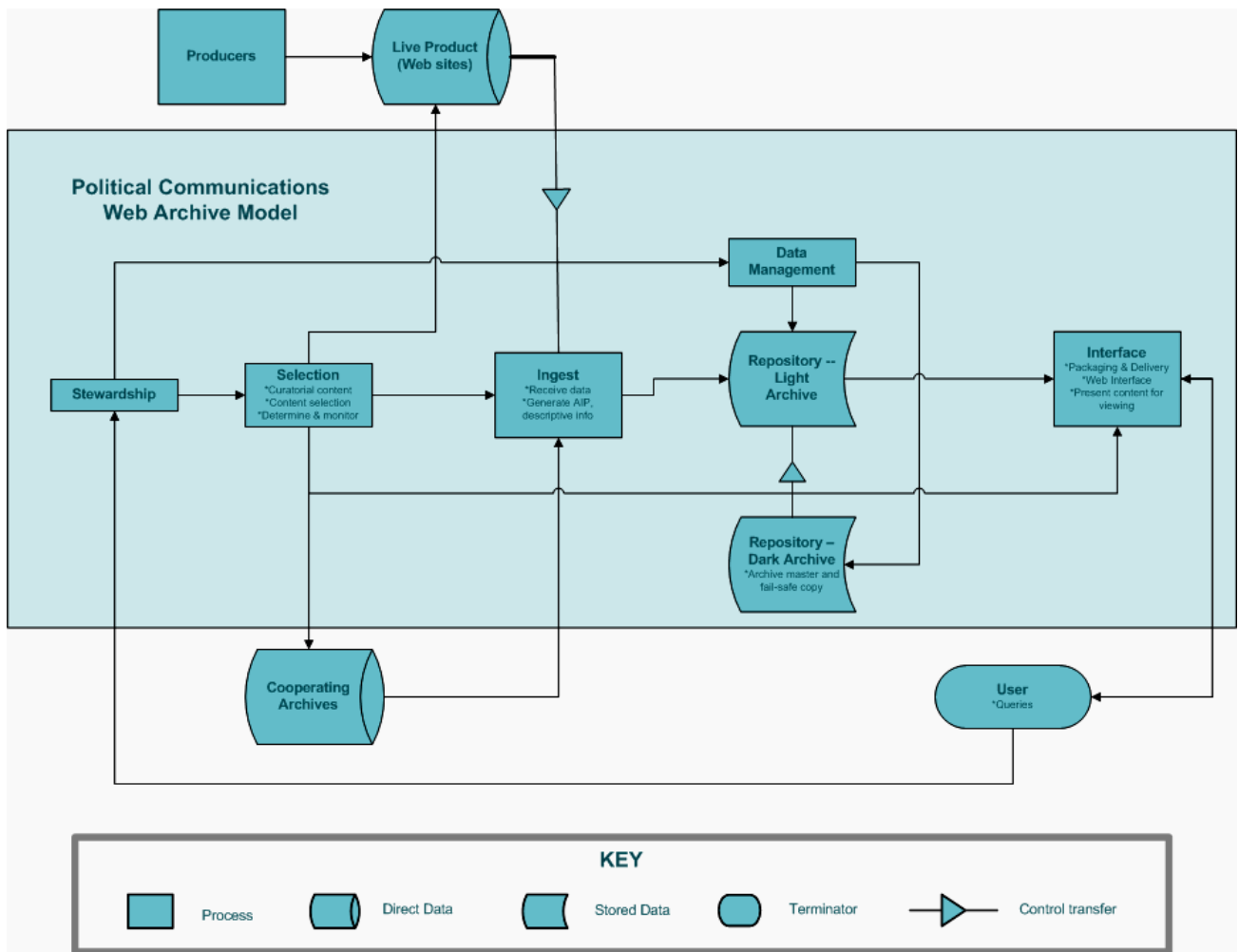
- o Ingest
- o Preservation planning
- o Data management
- o Archival storage
- o Administration
- o Access

The PCWA model also maps to the NDIIPP recommendations for architecture for long-term digital preservation. The NDIIPP architecture consists of four "layers," each of which performs a specific set of functions:

- o Repository
- o Gateway
- o Collection
- o Interface.

Recently Daniel Greenstein proposed a fifth layer, "Broker," to which the PCWA model "Stewardship" activity roughly corresponds. The term stewardship conveys a stronger sense of accountability to the user community than broker, which suggests an intermediary serving multiple parties and interests. Representation of the interests of the defined user community will be a critical function of the PCWA stewardship layer.

In accordance with the OAIS Model and NDIIPP architecture, the organization of activities or "layers" prescribed here permits a modular approach that allows the archiving infrastructure to be built in increments as additional resources become available and the "market" for archived materials matures. The proposed model also allows the Political Communications Web Archives to be built of services and components provided by different suppliers rather than by a single party, and permits upgrading and integration of new technologies over time. While the proposed model is a promising beginning, it will be necessary to refine it further to fully account for all of the necessary activities involved in the archiving of political Web materials.

Political Communications Web Archive Model

Producers

Live Product (Web sites)

Stewardship

Selection
*Curatorial content
*Content selection
*Determine & monitor

Ingest
*Receive data
*Generate AIP,
descriptive info

Data Management

Repository -- Light Archive

Repository – Dark Archive
*Archive master and fail-safe copy

Interface
*Packaging & Delivery
*Web Interface
*Present content for viewing

Cooperating Archives

User
*Queries

KEY

| Process | Direct Data | Stored Data | Terminator | Control transfer |

35

Service Model Activities or "Layers"

Selection / Curation:

Functional requirements: Selection/Curation activities involve identifying authoritative and appropriate content, and determining the technical and curatorial standards for capture and preservation of that content. Selection activities include:

- o "Prospecting," or searching the Web for political content on various topics, events, regions.

- o Identification of content to be captured and preserved. Selection specifications are expressed in terms of either specific or broad characteristics of sites. Specific characteristics might include, for instance, content under or linked to a single root URL or sites within a single domain[26]. General characteristics may specify content from a particular producer; produced in a specified language or dialect; produced in or pertaining to a specified region or country; sites that have specific functionalities or behaviors; or sites relating to specific events or topics.

- o Determination of the moment, frequency, depth, and scope of capture. This will involve assessing and addressing the risk of loss or disappearance of Web content, based on generalizations about the type of site, its subject, content, producer profile, and technical characteristics.

- o Identification of the "artifactual" characteristics of the target materials that must be documented. These characteristics might include, for example, the time/date of the instance captured, external links, document structure, URL, host, server type, authorship, authoring tools used, source code, metadata, etc, and would be documented in the AIP descriptive information.[27] Selection might also reach beyond the data attached in the site itself to domain registries, to capture information about the entity to which the site is registered.

- o Annotation of content and production of new descriptive metadata to ensure integrity and promote Administration of the content and its important evidentiary characteristics. Annotation can include language translation of source materials as well. [28]

Selection/curation activities will be most effective if distributed, that is, if they are able to take place at both local and centralized levels. "Local" Selection activities will populate archives developed and maintained by local or specialized communities, or even by individual researchers, to take advantage of specialized expertise that exists in specific communities of interest, such as a university Asian studies department, a foreign policy think tank, or an international relief agency.

Content valuable to the broader user communities will be included in a common central archive. Some content will be appropriate only for "dark archiving" and that determination will have to be made by selectors in accordance with standards provided by Management.

While most Selection activities will be undertaken by human selectors certain activities, such as determining the moment, frequency and scope of capture of a particular site or category of sites, might be pre-programmable and thus fully or partially automated. (As time passes and selection regimes are refined and codified a greater percentage of the Selection activities will be automated.) The archiving activities might involve a combination of automated and selective crawls, utilizing smart crawling

---

[26] Evidentiary characteristics are a more critical part of the "content" of such political evidence as statements, newspaper reports, government documents, and laws, more so than for electronic journals, where updating of information is critical. Characteristics like the number and nature of sites to which the target site links; number and nature of sites that link to target site, etc. can be important to establishing the credibility of the site in the absence of selector familiarity with site creators, content.

[27] Per *The Evidence in Hand* report from Council on Library and Information Resources, citing the need to ensure the research value, i.e., "originality, faithfulness, fixity, and stability–over time."

[28] This function can be separated from other Selection functions. The curatorial team suggested that Title VI National Research Centers "can recruit from the body of international students such programs tend to attract" for this kind of input activity.

techniques to assist in timing and selection, and incremental crawling to conserve bandwidth and storage costs.[29]

Selection should involve identification of current content for real-time archiving, and "retrospective" content for mining from existing archived materials, such as those harvested and archived by the Internet Archive, PANDORA, the Library of Congress Project Minerva and others, and perhaps materials retained in institutional or individual scholars' repositories.

Selection Behaviors: Automated or not, Selection must meet the requirements and interests of the local or specialized user communities as well as the users of a common PCWA archives. These requirements must inform the selection behaviors of specialists, who also generate content for the common archives as a secondary activity. Hence Selection for the archives should accommodate a variety of selecting behaviors:

- o *Project-based* - Individual scholars and researchers engaged in specific research projects at a university or policy institute.

- o *Ongoing/programmatic* - Selectors engaged in systematically building shared resources (e.g., Stanford's *Africa South of the Sahara* portal, LANIC, LC field offices and area studies departments) according to specified criteria governing a topic (immigration), genre (news) or region (Africa)

- o *Event-driven* - Creation of the Library of Congress Web archives *Election 2000* and *War in Iraq* responded to current events, and took place during or soon after those events.

The standards are shaped by the general content requirements for the common archives. Selection could take place in three ways:

1. *Original selection activities* -- Here Selection is performed by participating "accredited" specialists who identify and select content for the PCWA. These can be area and subject specialists assembling local e-archives at universities, libraries, policy institutes and perhaps Library of Congress field offices, or members of networks of scholars such as that being formed by Web archivist.org.

2. *Portal-based selection* -- Selection of PCWA content might also build upon portal-development and other, traditional political content-gathering activities. The effort might take advantage of identification, annotation, translation, indexing, and other work done by specialists and researchers in populating and maintaining region and subject portals, (like the LC *Portals to the World)*; Stanford librarians and specialists for *Africa South of the Sahara*; UN science and environmental policy experts for the *United Nations Environment Network* portal.

3. *Secondary selection* -- PCWA Selection might also draw content from certain broad nation- or domain- archiving and re-aggregating activities such as those of the National Library of Australia's PANDORA. Similarly, it might build upon the robust gathering and annotation activities supported by such content-sensitive organizations as the Foreign Broadcast Information Service.

Participants: Selectors would be members of a dynamic pool of agents authorized or "accredited" on the basis of standards determined and administered by Management (the users' proxy).

Accountability: Local users for local archive content, and the larger user community (Use / Consumption) through Management for Selection of common archives content. Selectors can be individual researchers, whose prospecting activities are driven by their own research agendas, and whose archiving activities are driven by their need to be able to "present" and source evidence through citation in their published works.

---

[29] Timing, for instance, might be triggered by a mechanism like D-Space's "digital provenance," that keeps track of changes in the digital object over time.

Requirements / Costs: High-level language, region, and historical expertise, transparency (of Selection criteria and funding); funding requirements will be a function of the volume of content targeted, the number and competency of participating selectors, the degree to which the frequency/timing regime must be customized (rather than standardized), the amount and complexity of annotation.   Costs for prospecting are difficult to determine, but will largely be absorbed by local support for professional or academic development of the Selector ("current awareness"). Costs overall here are also a function of the extent to which the activity is automated. In this activity this extent will probably remain low compared to those of Data Management, Ingest, and other activities that can be readily programmed. (Prospecting in particular is less likely to be automated.)

Incentives: Access to tools developed by central Data Management, and access to content archived in Repository selected by others. For individual researcher-selectors, the ability to present Political Web content as citations or evidence in their own published work.

Stewardship / Broker:

Functional requirements: Stewardship is a critical activity, upon which responsibility for continuous management of the archive's content and assets ultimately rests. Stewardship supports and monitors services and functions for the overall operation of the archiving effort; makes decisions and executes transactions pertaining to scope of the archives, participants, accessibility, and disposition of archives content and related assets; provides the nexus which gathers and pools the expertise and resources of diverse institutional and individual participants; establishes, formalizes, and monitors fulfillment of the terms for archiving activities, and ensures that standards and specifications for selection and presentation of content accord with User needs, as expressed in the Collection Development Policies and governance.

Stewardship functions include:

- o Securing submission arrangements with producers and other participating archives, and notification of hosts/producers on the means and terms of archiving (e.g., dark for 50 years);

- o Creating and administering policies regarding selection, access;

- o Authorizing or "accrediting" Selectors, Access providers, Administration, and other participants per standards established by Users or their proxies;

- o Monitoring and controlling Ingest of new content to the archives and Selector activity;

- o Certification or documentation of "authenticity" (chain of custody) of archives content.

Functions include financial asset management as well, since the scale of archiving activities will be reliant in part on the flow of funds and other resources. In this role Stewardship ensures that the scale of archiving activities and archives content are in line with the supporting resources. The addition of content to the archives, level of functionality in the presentation of content to users, and other cost-generating activities will have to correspond to levels of income or investment provided by the user communities directly or through their libraries and organizations.

Most important, Stewardship must incorporate and maintain governance mechanisms that ensure responsiveness of policy- and decision-making to the interests of the research community or Users. Such mechanisms might take the form of a governing council or Board of Directors consisting of representatives of the user communities.

Participants: To ensure availability and responsiveness of PCWA content to the larger user community, Stewardship activities should not be wholly or heavily reliant upon any one member or sector of the community, such as a single university, agency, or institute, without certain provisions. To the extent possible Stewardship should also be immune to constraints arising from individual local or national legal and political regimes, such as copyright, censorship, and other restrictions, that might compromise the

archive's inclusiveness. (Cite the FBIS problem.) Hence the Stewardship function would best be vested with a centralized entity, with mechanisms in place to make that entity accountable to the larger user/beneficiaries community, either directly or through authorized intermediaries (such as libraries, institutes, centers).

Stewardship activities should also be independent of Producers, to guarantee disinterested preservation of the artifactual integrity and the evidentiary characteristics of the content.

Peer-to-peer models for sharing of content that has been selected, archived, and stored by individual parties or communities can be adopted, and is possible using such tools as Herodotus, others. The usefulness of such materials to the larger community will depend on adherence to collective norms rather than the specific purposes of the individual project, agency or collecting organization under which it is archived.

Accountability / Funding: Users or their proxies. One governance mechanism proposed in curatorial team discussions was a governing board or advisory committee. To ensure sustainability and responsiveness the management activities must be undertaken by those accountable to the broad community of users/consumers, and not beholden to any single sector or faction of that community. This demands a funding model where most support derives from the user community. A mechanism might be provided by a subscription-based system for access or a similar funding structure that imposes a measure of User control over the archives Management activities.

Requirements / Costs: High-level legal, financial management expertise with regard to intellectual property rights, licensing, assert management; negotiation/brokering capabilities; trust; organizational stability and transparency; funding commensurate with the number of participants, users, and size of archives; scale and complexity of other archive activities. Stewardship activity must be undertaken by a legally constituted entity, such as a non-profit corporation, capable of entering into contracts for services and rights, holding and disposing of property, and accepting legal liability. The activity should be central to the mission of the Stewardship organization, and hence the organization should be neither a government entity (FBIS problem) or a single user party or its representative.

Ingest / Harvesting:

Functional requirements: Political Web content can be captured for archiving directly from the Web, or can opportunistically draw upon primary or "cooperating" archives assembled by other parties. In the first case Ingest activities include "pointing" or programming the Web crawler/harvester; undertaking the site crawls; receiving file data and capturing or generating the corresponding technical metadata, and notification of the Producer/Host about the archiving activities. This can be performed by an individual library selector or researcher in the process of a research project or local resource development effort, using standards and tools provided by the cooperative archiving effort. Ingest can also be undertaken on a larger scale by an organization or Administration (see below) agent as part of larger archiving activities. Under the right conditions Ingest might also harvest content from other, "cooperating archives." A great deal of Web content, much of it political, is harvested wholesale by organizations like Google, Alexa, and by various national efforts like that of the Australian National Library (PANDORA). Some of these "primary archives" are the products of federally funded domain-comprehensive archiving activities, such as the copyright deposit archiving in Denmark. Others are by-products of caching activities that have a commercial purpose, like those undertaken by Google and Alexa.

Ingest should be able to draw upon more specialized archiving activities undertaken in particular fields of interest by individuals or organizations with special area expertise, such as the Heidelberg University Chinaresource.org effort on Chinese Web materials, medical archiving undertaken by the Wellcome Institute, or the SSIC labor archiving activities.

In all cases Ingest crawls and saves Web sites and documents and replicates and locally stores the composite files according to Selection standards and specifications. Ingest involves:

- ensuring integrity of content and important metadata (the AIP descriptive information) per specifications determined by Selection and Administration.

- verifiably documenting, and annotating content with, circumstances of capture (such as date, method) providing baseline documentation for "chain of custody" of archived content.

Participants: For technical reasons Ingest activities may be inseparable from Data Management. Ingest could be performed by individual Selectors and researchers (using software like HTTrack) or by third-party Ingest/Harvesting agents. The latter could also include content-neutral commercial, government and non-profit service providers, like the Internet Archives, Google, and national libraries/legal deposit programs.

Accountability: Accountable to the Stewardship organization.

Requirements / Costs: High-level programming expertise and functionality are needed for centralized Ingest/Harvesting. High level expertise in file formats and the functionality of digital objects useful for Ingest at the local level. Costs are a function of the amount and structural complexity of the content targeted, and the degree of customization required to adapt crawls to target specific kinds of materials. (Comprehensive "indiscriminate" crawls entail lower programming costs.) The ratio of initial costs (again, programming) to ongoing costs is high.

Administration / Data Management:

Functional requirements: Administration activities involve monitoring and controlling data flows and auditing and certification of data and processes. Administration also monitors additions to the Repository, including content from other archives, to ensure that they meet appropriate technical standards, and provides feedback to Selection based on crawl results, changes in targeted materials, and changing factors in the crawl environment, to inform subsequent selecting activities and criteria.

Preserves functionality of archives content and migrates content to new platforms and formats as needed. Provides quality assurance of data by ensuring appropriate configuration and functionality of Ingest, Repository, and Access systems (hardware and software). Develops, procures and maintains tools and technologies for selection, annotation, and presentation of archives content, according to requirements established by Stewardship.

Interacts with Repository to provide system engineering requirements to monitor and improve archive operations and to inventory, report on, and migrate/update contents of the archive. Interacts with Access/Interface to ensure that Repository data is compatible with presentation functionality and determines the timetable under which to release archives content from dark to light repositories, according to terms established by Stewardship.

Interacts with Selection to ensure that archiving tools and technologies meet Selection needs. Administration is also responsible for implementing and monitoring adherence to archive policies and quality assurance standards regarding Ingest and Access, providing user support, and activating stored requests. *For technical reasons Data Management Activities may be inseparable from Ingest.*

Participants: Administration activities can be undertaken by content-neutral and user-neutral commercial or non-profit service providers. Some activity is automated.

Accountability: Accountable to Stewardship. To maximize Stewardship/User control, critical Administration functions cannot be reliant upon a single source of enabling software or technical platform. Activity is sensitive to technical functionality and formats of content, rather than to the subject or user inputs, which are input through specifications.

Requirements/costs: High-level analytical and technology expertise, robust network infrastructure and network management capability, high connectivity, high systems security. Costs are contingent on number of participants (selectors) and Repositories, and the volume and complexity of ingested and

stored content. High initial costs would be involved in establishing standards, methods, policies. Progressive automation of activity could yield savings here.

Repository / Secure Data Storage

Functional requirements: Serves as the "dumb" repository in the NDIIP architecture, where archived content would reside and be maintained per specifications established by Administration. The archives content would be made available to Access under the appropriate protocols and on a schedule and terms established by Administration, to permit temporary embargo of restricted content. Functions include:

- maintenance of bits

- storage of service content ("light archives")

- storage of comprehensive master content ("dark archives")

- storage of fail-safe content ("backup archives")

Accountability: Administration

Participants: Content-neutral commercial or non-profit service providers, such as the San Diego Supercomputer Center, OCLC, Internet Archive.

Requirements /Costs: High-level programming and analytical expertise, robust hardware infrastructure and network management capability, high-bandwidth connectivity, highest systems security. Costs are a function of the degree of rigor maintained in auditing and migration activities, volume and complexity of archived content; the number of discrete archives. There would be high initial costs which could be reduced by outsourcing most or all of Repository activity. With adequately precise specifications storage could be relegated to a contractor whose volume of work and specialized skills could result in savings and security. (This would add some cost for added specifications and quality control to Administration.) Costs should gradually decline, per historical trend in storage, offsetting at least partially the effects of growing content in the archives.

Access / Interface

Functional requirements: Access maintains the interface/delivery mechanisms that present light archives content for discovery and viewing by Users, and implements and maintains tools for User discovery and manipulation of archives content. Access activities can be undertaken either locally or centrally, as determined by the user community. Some archives Interfaces will be tailored to the needs of local or specialized constituents; others will be generic and designed to serve the needs of a wide range of Users.

Participants: Because Access / Interface activity is highly copyright-sensitive and content-sensitive, it may not permit participation by government or for-profit entities. Models include non-profit electronic publishers and presenters of research content, like the Research Libraries Group (Cultural Materials Initiative) and Library of Congress National Digital Library, others.

Accountability: Users and Stewardship.

Requirements / Costs: High-level expertise on User behaviors, and on the manipulation and presentation of research content. High-level programming expertise, robust hardware infrastructure and network management capability, high-bandwidth connectivity, high systems security. Costs are commensurate with complexity of content, level of functionality presented to the User and the degree of customization to distinct User communities, and (to a lesser extent) the number of authenticated users. Initial costs are relatively high while variable costs are low.

8. Next Steps

Because of the high cost of archiving Web materials and the relatively gradual pace at which Web materials are supplanting traditional primary source materials, Political Web archiving will have to be implemented incrementally. The higher education economy is now in contraction, and even in the best of times is relatively inelastic. The next stage of the PCWA effort will involve actualization of the framework specified in this report in a limited realm, as a "proof of concept." This approach will permit refining our understanding of the user needs and behaviors, testing of the proposed distribution of activities, and forming a more precise sense of the costs of each activity.

This might take one or more of the following paths:

o   Focused Real-Time Harvesting: Under the auspices of one or more of the Center's Area Microform Projects enlist a limited number of partners and, following the general curatorial and technical specifications outlined in this report, perform archiving over a one or two year period in a specific topic or domain.

o   Archiving Portal Materials: Collaborate with producers of a major region- or topic-based portal to build onto the portal effort an archiving component that provides persistent accessibility of sites and digital objects identified by the portal.

o   Retrospective Web Mining: Work with the Internet Archive and/or PANDORA to mine retrospective materials from their established archives and evaluate the suitability of those materials for research use.

From the economic and curatorial standpoints it may be best to focus during this stage on the capture and archiving of sites that are data-rich and have affinities with conventional news sites. This would build upon existing Center news archiving activities, such as the International Coalition on Newspapers, and would thus draw upon expertise that is already resident at or available to the Center. Such an approach would be likely to create synergies and economies with other Center area studies programs and with partner institutions.

If the Center is to undertake the Stewardship activities described in the proposed model, it will have to include representatives of additional communities of users. The policy research and international development communities are not now represented in the Center's membership or in the governance of its area studies programs. To ensure that the Political Communications Web Archiving effort is responsive to their needs and interests members of these communities will have to be "brought into the conversation" on shaping the program.

In the course of the project the Center for Research Libraries has established and strengthened a number of good and useful partnerships, with the Library of Congress Minerva Project, the Social Science Research Council, WebArchivist.org, and others, thus enlarging the circle of prospective participants in the next phase of the program. In the coming months the Center will be exploring the terms of engagement of potential existing and prospective partners. The next phase of the project might also draw upon the newly formed Ithaka organization to help it further develop the business model for the archiving effort.

The PCWA effort will likely contribute to and draw upon the continuing work of the California Digital Library, which is beginning to develop tools for selection and curation of Web materials generated by governments. The curatorial regimes and general technical requirements established in the PCWA investigation might help in shaping those tools. The tools developed, in turn, may be useful in subsequent PCWA archiving and curation efforts.

The outcome of the first Library of Congress NDIIPP awards competition will have an effect on which strategy the Center undertakes. There are several applications for archiving of Web sites, and the political Web effort might benefit from the outputs of one or more of those funded.

Attachment 1: User Survey

The project team mounted an on-line survey that targeted users of area studies Web content. (See the survey and results in Appendices 39 and 40). Users surveyed included those who accessed live Web materials, through portals like the Library of Congress *Portals to the World* and the University of Texas's Latin American Network Information Center portal, and those who studied archived materials, accessed through the Library of Congress on-line Web archive collections.

The purpose of the survey was to provide a sense of the potential PCWA users' needs, behaviors, and preferences that might shape the way Political Web materials were captured and archived. Some survey subjects were drawn to the survey through links on the Library of Congress Project Minerva Web site, the LANIC Web site, and Center for Research Libraries Political Web project site. Investigators solicited others through mailings to the Center's Area Studies Council and Area Microform Projects listservs and through mailings to investigation advisors and affiliated scholars. Respondents were informed that the purpose of the survey was to learn about the behaviors of researchers who used the Web as a primary or "citable" source of information. The survey was live for four weeks.

Most respondents were faculty (46.4%) and graduate students (30%) from the fields of History, Political Science, and International Relations, with a notable minority (19.5%) in Religion. The survey indicated that interest in materials from the Middle East and Northern Africa, U.S. and Canada, and Latin America and the Caribbean was high among respondents. Among the various kinds of sites accessed by the respondents the most frequently accessed were, in order of priority: news service, government, NGO, and protest or activist group sites. The domains considered most useful in their research were:

> .org (80.0%)
> .edu (75.2%)
> .com (56.8%)
> .gov (53.6%)

Most (84%) of the respondents said that they had accessed or "monitored" the same Web sites over time. Among sites monitored, however, the order of priority was slightly different than those simply accessed: news service, government, protest/activist groups, and then NGOs. While the subject regions of the researchers' interests were indexed in the survey, it is not possible from survey data to determine which percentage of the news sites accessed were produced in the subject regions and which, like the BBC, were Western or European sources.

Among sites which respondents themselves "archived," again news service and government sites were the highest. Respondents who archived site content did so most often by printing out, or by saving to local hard drives or servers. Fourteen of the 125 respondents (over 11%) said they had used the Internet Archive's *Wayback Machine™* to locate earlier versions of Web content; of these, eight indicated that it had been "somewhat useful." The majority of users (80%) indicated that an on-line archive of Political Web content would be useful in their field.

When asked what technical characteristics of sites were important to record or preserve, most respondents indicated that only the URL and date when accessed were necessary. Of the types of Web content captured by respondents who "archived" Web materials, a high percentage indicated text; fewer than 50% indicated images (although this may be attributable to the relative ease of archiving online text versus online images).

In an effort to determine the kinds of traditional materials that Web sources were supplanting for researchers, the survey queried what materials the researchers had used less frequently during the previous five years as citable sources. More than 10% of participants indicated that their use of newspapers (24.8%), government publications and documents (16.0%), journals (16.0%), or books (12.0%) had declined. Of those who answered the question, 67.6% indicated that such declines were due to increased use of Web-based resources, but 32.4% indicated that this was not the case.

Attachment 2: Studies of Individual Users

On November 18, 2003 the Library of Congress hosted a plenary meeting of the PCWA investigators and an assembly of scholars, library area studies specialists, and public policy researchers to review the preliminary findings and recommendations of the investigation. Investigating teams presented findings on the technical, curatorial, and organizational aspects of the study, and gathered feedback from those present that informed and shaped the conclusions in this report.

Investigators also conducted extended interviews with three different types of scholars engaged in advanced research for which Political Web materials were primary sources.   The interviews explored the nature of their research, the kinds of Web materials used, and the products generated by their work.   A summary of the findings follow.

Tomas Larsson.   Tomas Larsson is a Ph.D. candidate in Cornell University's Graduate School of Government. His dissertation topic is the evolution of property rights in land in Burma (now Myanmar) and Thailand after 1850. Larsson is examining the legal and administrative structures governing land ownership in Southeast Asia and how private and public property has been defined at various points in the region's history. He focuses on two periods in his research: the period 1890s through 1910, and the 1980s forward.

Larsson's chief sources of information on this project are:

> 1) Burmese (Myanmar) and Thai government Web sites, particularly those maintained by the departments of land and agriculture.

> 2) Burmese and Thai Political party Web sites, studied to determine the various parties' positions with respect to land issues.

> 3) On-line newspapers and newsgroup postings published in Thailand and Burma, and newspapers produced by exile communities. Larsson is most often alerted to news items usually via news compilation services, like the Burma Net News service (http://www.burmanet.org/) and other email notification or "clippings" services.

> 4) Sites maintained by foreign donors to the region, such as the World Bank and Australian aid organizations.

Larsson's discovery regime involves daily "grazing" of two newspapers (on-line versions), weekly scanning of three or four newspapers, and periodic consultation of several additional newspapers. Larsson also relies heavily on newsgroup information services provided by NGOs and other organizations, many of them non-profit. These groups push information to subscribers on selected subjects daily. He also searches Google frequently, using certain keywords to find new materials pertinent to his study.

Larsson noted the increasing reluctance of organizations and government agencies to make paper versions of their reports available when the same content is available on the Web. This reliance on Web delivery presents a problem when the reports are lengthy, and some are in excess of 400 pages. Larsson also notes that some materials from political groups and on-line magazines such as *Midnight University* are not available in paper form.

Larsson's use of hard copy Southeast Asian newspapers has become infrequent, because electronic versions of many titles are more readily accessible than the paper or microfilm versions. When he must cite content in his published work from on-line news Larsson will sometimes consult the hard copy version and reference that version in his citations. He believes that the paper editions do not have important information for his purposes that is lacking in the on-line versions. But Larsson cited as factors in his decision to cite the print source skepticism among colleagues about on-line sources and their uncertainty

about those sources' persistence. (He noted that BurmaNet archives its own bulletins on-line and features key word searching of them.)

Larsson archives important Web content for his personal use, employing EndNote, a bibliographic management software produced by Thomson ISI. EndNote allows researchers to create their own personal database of references to articles, books and other collected materials, and works in tandem with word-processing software. Larsson saves important Web content, usually single pages or documents, to his local hard drive in its original form (usually HTML or PDF), and then links those documents to his bibliography in EndNote. (EndNote is proprietary software and not a true archiving solution.)

The information about the sites (metadata) that he collects includes the URL of the home page or the document of interest and the time and date of capture. The URL is normally sufficient to indicate to him the source of the materials or the identity of the producing organization. The structure of the site and the linkages between the pages are not of interest to Larsson beyond their use in being able to maintain the integrity of the selected texts.

Larsson considers textual materials from the sites far more important to his research than images, which he encounters infrequently. He also professes a willingness to rely on and cite in his research texts reported in digests and compilations, rather than in the original sources. Larsson suspects, however, that these compilations do not always get the original publication dates correct.

****

Priscilla Offenhauer. Dr. Offenhauer is a historian by training, with a Ph.D. in European Intellectual and Social History, and is a Research Analyst in the Library of Congress Federal Research Division. Offenhauer recently contributed to a multi-year ongoing research project on the trafficking of women and children across international boundaries for illegal purposes, undertaken for cooperating government agencies. The ultimate goal of the project is a comprehensive on-line database and archive of information about human trafficking worldwide, a database that can provide the law enforcement, public policy and NGO communities ready access to sourced, trustworthy information on the magnitude and geographical distribution of trafficking activity.

Offenhauer's role was to gather and compile sourced materials from "gray literature" e.g., conference papers, think tank and government reports, government policies and position papers, substantial news bulletins, statistical and narrative reports, and other materials; to annotate them regarding source and timing of issuance; and to provide them for inclusion in the database. Since the material gathered and provided was to support further action by governments and NGOs the credibility of the content was a large factor. Hence careful sourcing and preservation of the evidentiary traits of the material were important to Offenhauer's work.

Offenhauer focused on source materials produced since the year 2000. Many of these materials she extracted from local government and NGO Web sites in the former Soviet Union, Eastern Europe, Southern and Southeast Asia, and Africa, and elsewhere where trafficking is active. Offenhauer also captured postings from well-monitored newsgroups devoted to trafficking, such as Stop-traffic, and news services like the Foreign Broadcast Information Service. Offenhauer monitored trustworthy major sites as well, like that of the International Organization for Migration (http://www.iom.int/) and the United Nations. Newspapers were not a primary source for her work, however, because the accounts of trafficking they contained -- usually from police reports -- tended to be accounts of individual incidents and secondary accounts of study results. But news did provide a means of learning second-hand about reports that synthesized information about such incidents into more broadly descriptive sources, such as newly published reports from the United Nations or the International Organization for Migration. Offenhauer also relied on notification through listservs and discussion groups to alert her about such reports.

Offenhauer noted that many of these materials were available in print but were far easier to discover and obtain on-line than they had been when in the past they were only available on paper. This was true for a

number of reasons. In most libraries gray literature is usually given low priority for cataloging, or is maintained "off-line" in file cabinets. Books, on the other hand, are published and made available too slowly to stay current with the topic under consideration, where developments unfold quickly and policy is based on fresh information.   In Offenhauer's own words:

> "The materials I sought were substantial reports (from international organizations, governments, NGOs, and academe), which would have paper versions. However, unless I could get them directly from the authors (as I did in some cases), the surest way to get them was from web postings. Using the web, I avoided bottlenecks in library processing, and in ordering materials from publications offices (of, say, the U.N.) The percentage of materials I collected from the web was perhaps 75."

However, the fugitivity of these materials was cited by Offenhauer as a serious concern: many of the Web sources Offenhauer was citing had already disappeared during the course of her project. This issue was serious because of the importance placed on "sourcing" materials for the end product. Early in her project Offenhauer devised ways to print out full texts and statistical materials from the Web, and to "cut and paste" this electronic content into word-processing files for future reference and citation.   For purposes of sourcing the materials she archived Offenhauer considered it sufficient to capture the URL of the site, the date accessed, the name of the producing organization, and the relevant textual content. This information was enough to establish the necessary context and authenticity for her purposes. In some instances she also captured images where they existed.

****

W. Sean McLaughlin.   McLaughlin is a senior analyst with DFI Government Services, a Washington DC-based defense consulting firm specializing in homeland security issues. McLaughlin was chosen for an interview because of his published research analysis of changes in Political Web materials over time. Unlike the other researchers interviewed, for McLaughlin the medium itself was the message. Where other researchers mined the contents of Political Web communications for information on actual events, McLaughlin studied the communications strategies adopted by selected political actors, analyzing the changes in those strategies and the messages they conveyed during a finite period.

McLaughlin's "The Use of the Internet for Political Action by Non-state Dissident Actors in the Middle East," published in *First Monday* in November 2003, is a lengthy and revealing case study of Political Web production.[30] Research for the publication was undertaken for a senior honors thesis at Georgetown University under the direction of Professor Bernard I. Finel, the executive director of the university's M.A. in Security Studies Program and the Center for Peace and Security Studies.

McLaughlin studied multiple successive instances of more than two dozen Web sites maintained by three dissident groups: the Muslim Brotherhood in Jordan, Muslim Brotherhood in Egypt, and the Movement for Islamic Reform in Arabia. His research involved monitoring changes in the sites produced by the subject organizations by accessing those sites, at weekly intervals, throughout 2001-2002.

McLaughlin's study provided a great deal of information about the behaviors of Political Web producers, particularly about the activities of dissident groups in a region where censorship and other state-imposed constraints disrupt traditional channels of communication between the groups and their supporters. He showed, for instance, how the Movement for Islamic Reform in Arabia (MIRA), founded in 1996 to promote Islamic reform within the Saudi kingdom, crafted its use of Web communications to elude detection and to accommodate the horizontal, non-hierarchical structure of this trans-national organization.

McLaughlin supplemented his real-time monitoring of the sites with the Internet Archive's Wayback Machine. The Wayback Machine was a source of comparative material, namely of past instances of some of the subject organizations' sites from as early as 1996. He also used the Wayback Machine in his

---

[30] Reference: http://www.firstmonday.org/issues/issue8_11/

published article as a reliable source for making viewable for reference citations to subject sites that had since disappeared.

McLaughlin sees the archives available through the Wayback Machine as vital to his study but somewhat limited. He indicated that certain kinds of site content that the Wayback Machine did not preserve, such as images, captions, and sound and multimedia files, might have been useful in his study.   McLaughlin noted that a great deal of multimedia content, like Arabic language audio recordings available on some of the sites had been lost, He remarked that the Saudi dissident groups rely heavily on recorded messages, some as long as thirty minutes, which change frequently, even weekly. Yet despite the occasional losses of visual and audio content, however, McLaughlin felt he was able to get "an accurate picture of the political environment."

<center>BIBLIOGRAPHY</center>

On-line Politics and Journalism

Beyerla, Shaazka. "The Middle East's e-War." _Foreign Policy_, July-August 2002.

Boyd., Andrew "The Web rewires the movement: grassroots organizing power of the Net." _The Nation,_ August 4, 2003.

"Dusting off the Search Engine; the history of indexes of the New York Times." New York Times, November 17, 2001.

Getler, Michael. "Caught in the Crossfire: media coverage of tie latest Palestinian uprising." Washington Post, May 5, 2002.

Gray, Louise. "Protest Web Sites." _The New Internationalist_, July 2003.

"Iran: minister identifies 170 'counterrevolutionary and political' web sites." _Asia Africa Intelligence Wire_, September 22, 2003

Jarvis, Michael . "Net Effect: Spinning History. Political Web sites_. Foreign Policy_ Jan-Feb 2003.

Kalathil, Shanthi and Taylor C. Boas. "The Internet and State Control in Authoritarian Regimes: China, Cuba and the Counterrevolution." _First Monday_, August 2001. http://www.firstmonday.org/issues/issue6_8/kalathil/index.html

Latham, Robert ed., _Bombs and Bandwidth: the Emerging Relationship between Information Technology and Security_. New York and London: The New Press, 2003

"Liberia Independent newspaper closed: The Analyst." _New York Times_, April 26, 2002

McLaughlin, Sean, "The use of the Internet for political action by non-state dissident actors in the Middle East" First Monday, November 3, 2003. http://www.firstmonday.org/issues/issue8_11/

Mark, David . "Four Informative Political News Web Sites." _Campaigns and Elections_, February 2003.

----------- . "Legislative Web Sites as Campaign Tools." _Campaigns and Elections_, Sept ember, 2002Rohozinski, Rafal, "Bullets to Bytes: Reflections of ICTs and 'Local' Conflict" in Robert Latham, ed., _Bombs and Bandwidth: the Emerging Relationship between Information Technology and Security_. New York and London: The New Press, 2003.

Trofimov, Yaroslav , "Arab opinion softens amid Afghan blitz: in widely read newspapers, a new self-criticism; anti-U.S. news eases a bit." _Wall Street Journal_, November 26, 2001

Williamson, Andy. "The Impact of the Internet on the Politics of Cuba." First Monday, August 2000 http://www.firstmonday.org/issues/issue5_8/williamson/index.html

Also:

The "Net Effect" column of _Foreign Policy_ magazine, published by the Carnegie Endowment for World Peace) provides good topical digests of selected Web sites in various parts of the world.
See also the excellent periodic coverage of the Political Web and on-line journalism by Michael Getler in the Washington Post, and Felicity Barringer in the New York Times.

Curatorship and Technology

Arms, William Y. (2001) "Web Preservation Project Final Report." A Report to the Library of Congress http://www.loc.gov/minerva/webpref.pdf

Arvidson, Allan, Krister Persson, and Johan Mannerheim (2000) "The Kulturarw3 Project-The Royal Swedish Web Archiw3e-An Example of "complete" collection of web pages." A paper presented at *the 66th IFLA Council and General Conference*, Jerusalem http://www.ifla.org/IV/ifla66/papers/154-157e.htm

Besek, June M. (2003) "Copyright Issues Relevant to the Creation of a Digital Archive: A Preliminary Assessment." Council on Library and Information Resources and Library of Congress, Washington, D.C http://www.clir.org/pubs/reports/pub112/pub112.pdf

Cedars (n.d.) "Metadata for Digital Preservation: The Cedars Project Outline" http://www.leeds.ac.uk/cedars/colman/metadata/metadataspec.html (accessed 12.10.2002)

Chapman, Stephen "Counting the Costs of Digital Preservation: Is Repository Storage Affordable?" http://jodi.ecs.soton.ac.uk/Articles/v04/ i02/Chapman/chapman-final.pdf

Charlesworth, Andrew (2003) "Legal issues relating to the archiving of Internet resources in the UK, EU, USA and Australia." A study undertaken for the JISC and the Wellcome Trust http://library.wellcome.ac.uk/projects/archiving_legal.pdf

Christensen-Dalsgaard, Birte, Eva Fønss -Jørgensen, Harald von Hielmcrone, Niels Ole Finnemann, Niels Brügger, Birgit Henriksen, and Søren Vejrup Carlsen (2003) "Experiences and Conclusions from a Pilot Study: Web Archiving of the District and County Elections 2001." *Final Report for The Pilot Project "netarkivet.dk"* http://www.netarkivet.dk/rap/webark-final-rapport-2003.pdf

Council on Library and Information Resources (2000) "Authenticity in a Digital Environment" http://www.clir.org/pubs/reports/pub92/contents.html

Day, Michael (2003) "Collecting and preserving the World Wide Web" A feasibility study undertaken for the JISC and Wellcome Trust http://library.wellcome.ac.uk/projects/archiving_reports.shtml

DELOS/NSF Working Group (2003) "Reference Models for Digital Libraries: Actors and Roles." *Final Report* http://www.delos-nsf.actorswg.cdlib.org/

Gladney, Henry M. (1999) "Digital Dilemma: Intellectual Property." *D-Lib Magazine*, Vol. 5, No. 6, December http://www.dlib.org/dlib/december99/12gladney.html

Gross, Jennifer (2003) "Learning by doing: The Digital Archive for Chinese Studies (DACHS)." Paper given at the *3rd ECDL Workshop on Web Archives* http://www.sino.uni-heidelberg.de/dachs/publ.htm

Hodge, Gail M. (2000) "Best Practices for Digital Archiving: An Information Life Cycle Approach." *D-Lib Magazine*, Vol. 6, No. 1, January http://www.dlib.org/dlib/january00/01hodge.html

Institute of Chinese Studies, University of Heidelberg (2002) "About DACHS: Introduction" http://www.sino.uni-heidelberg.de/dachs/.intro.htm (accessed 12.10.2002)

Kenney, Anne R., Nancy Y. McGovern, Peter Botticelli, Richard Entlich, Carl Lagoze, and Sandra Payette (2002) "Preservation Risk Management for Web Resources: Virtual Remote Control in Cornell's Project Prism." *D-Lib Magazine*, Vol. 8, No. 1, January http://www.dlib.org/dlib/january 02/kenney/01kenney.html and http://www.dlib.org/dlib/january 02/kenney/kenney-notes.html

Koman, Richard (2002) "How the Wayback Machine Works" http://www.oreillynet.com/lpt/a/1295 (accessed 12.10.2002)

Library of Congress (2003) "METS: An Overview & Tutorial." *Metadata Encoding & Transmission Standard (METS) Official Web Site* http://www.loc.gov/standards/mets/METSOverview.html (accessed 06.26.2003)
—— "MODS." *Metadata Object Description Schema Official Web Site* http://www.loc.gov/standards/mods/ (accessed 06.26.2003)

Lyman, Peter (2002) "Archiving the World Wide Web." In *Buillding a National Strategy for Digital Preservation: Issues in Digital Media Archiving* Council on Library and Information Resources and Library of Congress, Washington D.C. http://www.clir.org/pubs/reports/pub106/web.html (accessed 10.16.2003)

Masanès, Julien (2002) "Towards Continuous Web Archiving: First Results and an Agenda for the Future." *D-Lib Magazine*, Vol. 8, No. 12, December http://www.dlib.org/dlib/december02/masanes/12masanes.html

NINCH (2002) "Digitization and Encoding of Text." In *The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials* http://www.nyu.edu/its/humanities/ninchguide/V/ (accessed 03.10.2003)
—— "Digital Asset Management." In *The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials* http://www.nyu.edu/its/humanities/ninchguide/XIII/ (accessed 03.10.2003)
—— "Preservation." In *The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials* http://www.nyu.edu/its/humanities/ninchguide/XIV/ (accessed 03.10.2003)
—— "Appendix A: Equipment." In *The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials* http://www.nyu.edu/its/humanities/ninchguide/appendices/equipment.html (accessed 03.10.2003)
—— "Appendix B: Metadata." In *The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials* http://www.nyu.edu/its/humanities/ninchguide/appendices/metadata.html (accessed 03.10.2003)
—— "Appendix C: Digital Data Capture." In *The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials* http://www.nyu.edu/its/humanities/ninchguide/appendices/capture.html (accessed 03.10.2003)

OCLC/RLG Working Group on Preservation Metadata (2002) "Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects." OCLC Online Computer Library, Inc., Dublin, Ohio.

Oracle Technology Network (2002) "Ultra Search Crawler Extensibility API" http://otn.oracle.com/products/ultrasearch/index.html

Pandora Archive (2003) "Documents and Manuals" http://pandora.nla.gov.au/documents.html
—— "Online Australian Publications: Selection Guidelines for Archiving and Preservation by the National Library of Australia" http://pandora.nla.gov.au/selectionguidelines.html

Research Libraries Group (2002) "Trusted Digital Repositories: Attributes and Responsibilities." An RLG-OCLC Report http://www.rlg.org/longterm/repositories.pdf

Russell, Kelly, and Ellis Weinberger (2000) "Cost elements of digital preservation" http://www.leeds.ac.uk/cedars/colman/CIW01r.html (accessed 12.10.2002)

Seville, Catherine, and Ellis Weinberger (2000) "Intellectual Property Rights lessons from the CEDARS project for Digital Preservation" http://www.leeds.ac.uk/cedars/colman/CIW03.pdf

W3c, "Web Characterization Terminology & Definition Sheet," http://www.w3.org/1999/05/WCA-terms/

Weinberger, Ellis (2000) "Towards Collection Management Guidance" http://www.leeds.ac.uk/cedars/colman/CIW02r.html (accessed 12.10.2002)

Wimmer, Walter (2002) "Automized production of bibliographical information of locally stored Internet files: a project to establish archives of electronical press services of parties and trade unions." In *Acta/International Association of Labor History Institutions* http://library.fes.de/fulltext/bibliothek/01103.htm