

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

Introduction: Aims and Background of the Investigation	1
Participants and Advisors to the Project	3
1. The Political Web - Production and Producer Behaviors	5
2. User Behaviors and Needs	7
3. Existing Approaches to Archiving Traditional and Web-based Political Materials	9
4. Curatorial Regimes and Issues	14
5. Technical Strategies	21
6. Sustainable Archiving - How Best to Organize, Govern, and Fund the Activities	25
7. A Proposed Political Communications Web Archives Model	32
8. Next Steps	42
Bibliography	50
Appendices	

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 public 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 monographs² 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 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 gifs, 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 Heritrix⁸

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 Kulturaw3 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 Kulturaw3 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: .org, .net and .nu. Kulturaw3 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 identified 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 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 characteristics 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 characteristics 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 factors 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 archiving 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/masanés/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 model¹⁹ and the Trusted Digital Repository model²⁰ 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://www.classic.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. HTRACK/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 which 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

- Prospecting for archives content

²² A second commercial pilot case is Netflix, an online DVD rental service which allows browsing and 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

- Presenting content for viewing and searching
- 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 Kulturaw3 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 a 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 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:

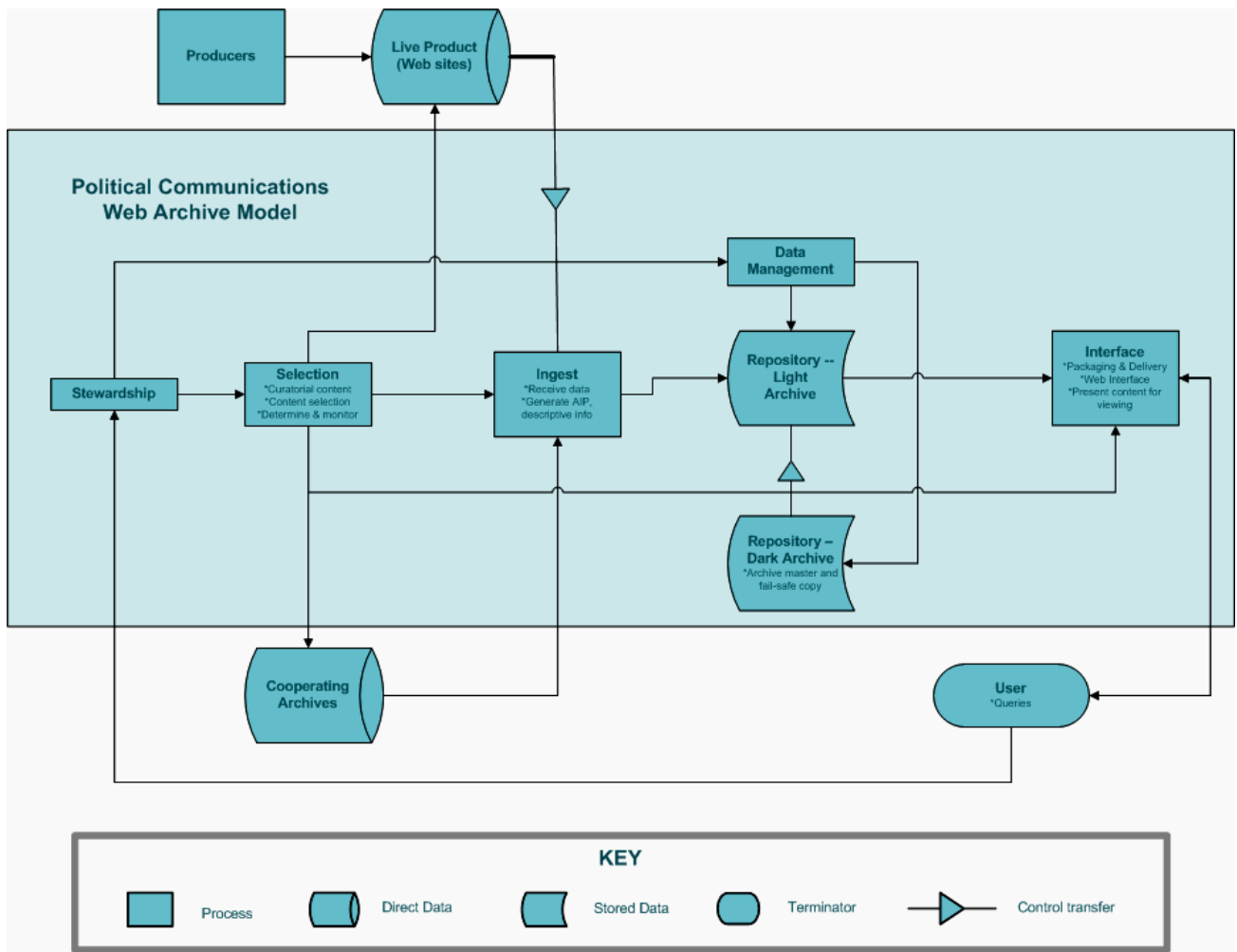
- Ingest
- Preservation planning
- Data management
- Archival storage
- Administration
- Access

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:

- Repository
- Gateway
- Collection
- 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.



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:

- “Prospecting,” or searching the Web for political content on various topics, events, regions.
- 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 domain²⁶. 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.
- 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.
- 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.
- 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:

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

- 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);
- Creating and administering policies regarding selection, access;
- Authorizing or “accrediting” Selectors, Access providers, Administration, and other participants per standards established by Users or their proxies;
- Monitoring and controlling Ingest of new content to the archives and Selector activity;
- 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 NDIIIP 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:

- 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.
- 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.
- 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 Offenbauer. Dr. Offenbauer 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. Offenbauer 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.

Offenbauer'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 Offenbauer's work.

Offenbauer 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. Offenbauer also captured postings from well-monitored newsgroups devoted to trafficking, such as Stop-traffic, and news services like the Foreign Broadcast Information Service. Offenbauer 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. Offenbauer also relied on notification through listservs and discussion groups to alert her about such reports.

Offenbauer 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 Offenbauer'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 Offenbauer as a serious concern: many of the Web sources Offenbauer 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 Offenbauer 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 Offenbauer 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.”

BIBLIOGRAPHY

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 the 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*, September, 2002
Rohozinski, 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 *Building 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/masanés/12masanés.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>

APPENDIX 1

Archival Access Policy survey

Adrienne Sonder

June 19, 2003

The material below is excerpted and/or copied from the sources. Information was gathered from U.N.-affiliated organizations, and U.S. federal and state agencies. Findings suggest that few records at the state level have prescribed closed periods. The number is also small at the federal level. In international archives, the closed period varies depending on the organization and type of record. The shortest close period, imposed by the U. N., restricts records for a period of 20 years if the records were closed upon acquisition or if the confidentiality of the record is in question.

International organizations:

Selected guidelines for the management of records and archives : A RAMP reader. Prepared by Peter Walne for the General Information Programme and UNISIST. - Paris, Unesco, 1990.

Retrieved from:

<http://www.unesco.org/webworld/ramp/html/r9006e/r9006e0q.htm#Access%20to%20the%20archives%20of%20united%20nations%20agencies>

[Ch. 25- Access to the archives of United Nations agencies]

6.2 The State of Affairs

Records/archives lacking standards and procedures for classification and declassification, retention periods, disposal policies and realistic conditions of access mean frustration to archivists as well as to internal and external users. The present survey has revealed a number of inadequacies in regard to the international organisations. Among the sample of the 34 international organisations chosen, only 41.2 per cent answered the questionnaire in a comprehensive manner. What is happening, if anything, in the field of archives administration in the other 58.8 per cent? So much information is missing that it seems almost impossible to get a clear picture of the actual situation...

..In general, rules and procedures relating to archives are rather scarce in the international organisations, although these instructions are essential in the maintenance of an operative records/archives management service. In this survey 11 organisations reported that they have such instructions, but only five submitted the texts as requested. Instructions of the IMF, UN Secretariat (also followed by UNECA and UNOG), UNESCO, UNICEF and WHO satisfy the standards of what is considered to be good archives administration. Otherwise, the so-called instructions are simply correspondence and registry manuals for secretaries, if the organisation has even such instructions....

6.3 Diversity in Access

Accepting the definition of access as "the availability of records/archives for consultation as a result both of legal authorisation and the existence of finding aids" means detailed responsibilities for archives administration. The manner in which UN agencies are dealing with this question differ in many respects and, for that reason, it is of interest to examine the content of selected rules and procedures.

6.3.1 The United Nations Archives

An Administrative Instruction, ST/AI/326, of 28 December 1984 "explains the guidelines concerning internal and public access to the United Nations archives". Access is given both to archives and non-current records kept by the service. It is clearly stated that staff members of the Secretariat may have access if they need the documents for official business, "except those subject to restrictions imposed by the Secretary-General". Regarding public access to archives and records, it is asserted that:

- (a) they are open if they were accessible when created;
- (b) they are open if they are more than 20 years and not subject to restrictions; and,
- (c) they are open if they are less than 20 years and not subject to restrictions.

Consequently, the United Nations Secretariat follows a time limit of 20 years, but with flexibility in regard to non-restricted material. With respect to restricted records the Secretary-General has imposed two levels of classification:

- ST - Strictly Confidential to records originating with the Secretary-General, the unauthorised disclosure of which could "cause grave damage to confidence in the Secretary-General's Office(s) or to the United Nations".

- SG - Confidential to records originating with the Secretary-General, the unauthorised disclosure of which could "cause damage to the proper functioning of the United Nations Secretariat".

"SG - Confidential" records are automatically declassified when 20 years old, and "SG - Strictly Confidential" are reviewed for declassification at this age. Declassification in either case can be approved prior to the expiration of 20 years.

The United Nations Archives rule of a 20 year time limit is gaining wider acceptance, as in the case of UNESCO and UNICEF, and it could be a starting point in discussions on the subject of access.

6.3.2 UNESCO Archives

The "Rules governing access by outside persons to UNESCO's Archives" reveal that the holdings consist of documents, field mission reports and records. The first two are "freely accessible in the reading room of the Archives Section", although documents can be marked "restricted and confidential" and access given only "if the prior agreement of the relevant unit of the Secretariat has been obtained". Often, the documents are mimeographed or other multicopied material but not archival documents.

The third category, records, is another case. According to the Chief Archivist, a relaxation in access is currently under consideration, following the UN Secretariat's rule of 20 year time limit. Until any changes are made, the rules in force place it at 30 years, "with the exception of certain types of material where UNESCO may decide on a shorter period". A closed period limit of 50 years is imposed on the following material:

- files containing exceptionally sensitive information on relations between Unesco and its Member States, between Unesco and the United Nations, intergovernmental and non-governmental organisations;
- files containing papers which, if divulged, might injure the reputation, affect the privacy or endanger the safety of individuals;
- personnel files of officials or agents of Unesco; and,
- confidential files of the offices of the Unesco Director-General; Deputy Director-General and Assistant Directors-General.

It should be stressed that access to archives within the open period can be refused if they are "unmistakably of confidential nature still" and exceptions "to a paper or file that is not yet in the open period may be made by the Chief Archivist" after some provisions are fulfilled. The UNESCO rules thus also have a degree of flexibility.

6.3.3 UNICEF Records and Archives

This organisation has adopted rules and regulations similar to those promulgated by the UN Archives. The "Procedural guidelines for UNICEF records and archives" of 9 November 1983 follow closely the access conditions and 20-year rule adopted by the UN Secretariat. Archives and non-current records follow the same pattern of consultations and restrictions. Except that the latter can be imposed either by the Secretary-General of the UN, the Executive Director of UNICEF or their authorised representatives.

6.3.4 WHO Archives

These archives are defined primarily as "documents and correspondence of various kinds, received or produced by the Organisation in the course of carrying out its functions, and which have been

preserved in whatsoever form for documentary and historical purposes. External material, whether public or private, relating to the activities of the Organisation may be added to the archives; such material shall also be subject to these rules". That reference appears in "Rules governing access to WHO Archives" of 15 February 1974.

Access is given in situ after a time limit of 40 years but more recent material can also be freely consulted if it does not have any confidential component. In practice a pragmatic 10-year time limit is also employed. The determination of what is confidential is a prerogative of the organisation and is not clarified in the rules. WHO Archives also has material with closed periods of up to 60 years, i.e. "files containing information which, if disclosed, might prejudice the reputation, personal safety or privacy of individuals".

6.3.5 IMF Archives

This organisation applies no time limit for access to its holdings. "General Administrative Order No.26, Rev.I" of 1st November 1969, states: "All Fund documents and other records shall be considered restricted and not for public use except when designed for transmission to the public or specifically authorised for distribution to a particular recipient or group of recipients". The documents may also be classified as confidential or secret:

- "Confidential - records containing information, the unauthorised disclosure of which might be prejudicial to the interest of the Fund or its members. Records, the subject of which required limitations on use for reason of administrative privacy.
- Secret - records containing information, the unauthorised disclosure of which would endanger the effectiveness of a program or policy, or hamper negotiations in progress, or which could be used to private advantage. Use of this classification should be held to an absolute minimum".

6.3.6 Overview

In summary, from the above examples, it appears that access to the records/archives of international organisations is related to the identification of what is in the archives: the interpretation of the right to information;; respect for privacy of individuals; and the protection of the organisation's different spheres of interest. In addition, to open archives to the public means that the organisation must comply with basic requisites, including a good record management system and the provision of user facilities. These goals have not been realised in many international organisations at the present time...

Guide to the archives of intergovernmental organizations. International Council on Archives. (Not dated). Retrieved online June 19, 2003 from:
<http://www.unesco.org/archives/guide/uk/sommaire2.html>

** This site lists a number of international organizations with information on their archives administration policies. Listed below is a selection of those organizations that specified actual time periods to keep records sealed.*

1. International Federation of Red Cross and Red Crescent Societies

Access rules

The archival records are open to the public by appointment with the archivist, and in accordance with the following access conditions:

- i. The Secretariat classifies as public the following records:
 - a. Federation publications that the Secretariat makes available for sale to the public or distributes to the public for free;
 - b. Decisions of the General Assembly, and policies or reports adopted through a Decision, except for those decisions, policies and reports designated confidential by the General Assembly.
 - c. Decisions of the Governing Board, and reports adopted through a Decision, except for those decisions and reports designated confidential by the Governing Board.
 - d. minutes and reports of statutory bodies more than 20 years old;

- e. non-confidential files of the Secretariat that are more than 30 years old.
- ii. The period after which a record becomes public is calculated from the date on which the record is closed.
- iii. Records classified confidential, which are generally records containing personal data, are closed to the public.

2. The Food and Agriculture Organization of the United Nations (FAO)

Access Rules

The archival records of FAO are available for consultation by staff members in the course of their official duties *in situ* or on loan. Non-staff members, demonstrating a legitimate interest, may be given access to archival records, which have been closed for 15 or more years. In addition to the general 15 years closure period of records, special restrictions apply to records of confidential nature, such as personnel files of separated staff members, confidential reports, etc. The Director-General may, in appropriate circumstances and on recommendation of the Chief, Records and Archives Unit, remove these restrictions.

3. International Committee of the Red Cross (ICRC)

Access Rules

In January 1996, the ICRC Assembly adopted new "Rules governing access to the archives of the ICRC", which gave the public unrestricted access to archives dating from before 1950.

This historic decision was taken to respond to the desire of historians and many other people in search of accounts regarding individual victims of conflict and the conflicts themselves to extend the historical research undertaken since the late 1970s, at the initiative of the ICRC itself. A noteworthy example of this is a book entitled *Une mission impossible? Le CICR, les déportations et les camps de concentration nazis*, written by Professor Jean-Claude Favez of the University of Geneva, with initial publication in 1988 and a new edition appearing in 1996.

Extract from the "Rules governing access to the archives of the ICRC" of 17 January 1996.

"SECTION III: PUBLIC

Public archives Art. 6 : The general public has access to archives classified as public. The ICRC archivists select and make an inventory of archives to be classified as "public". After a set period of time, to ensure that such access will in no way be detrimental to the ICRC, to the victims that it is its duty to protect, or to any other private or public interests requiring protection.

Public archives Art. 7 :

¹⁾Three types of document are to be found in the "public" archives :

- General ICRC files dating back more than 50 years, including minutes of the decision-making bodies;
- The minutes of the Recruitment Commission, the personal files of staff members and the record series containing personal or medical information dating back more than 100 years :
 - Access to biographical or autobiographical information on a specific individual is allowed after 50 years; such research, however, must be carried out by an ICRC archivist (see Article 10);
 - If permission is obtained from the individual concerned, the 50-year period may be shortened;
- Access to archival material from other sources which has been stored in the ICRC archives is authorized from the date set by the individuals or institutions that deposited the material at the ICRC.

²⁾The period during which the public is barred from consulting a file runs from the date on which the file is closed.

³⁾Documents that were open to consultation by the general public before being deposited in the ICRC archives remain so thereafter.

Special access Art. 8 :

¹)The Executive Board may, before expiry of the time limits set in Article 7, grant special access to facilitate academic work which the ICRC itself wishes to see successfully completed or which it finds of interest.

²)The Executive Board adopts the *Rules governing special access to the ICRC's classified archives. Restrictions* Art. 9 : Public access to ICRC archives may be temporarily delayed in order to permit necessary conservation work to be carried out on the documents requested, or if no space is available in the reading room.

Fees Art. 10 : A charge is made for research carried out by ICRC staff at the request of persons outside the organization (see Article 7).

Use Art. 11 : No use may be made of the archives for commercial purposes unless a specific contract to that effect has been concluded with the ICRC."

With regard to access to the ICRC archives, see also Jean-François Pitteloud, "New access rules open the archives of the International Committee of the Red Cross to historical research and to the general public", in *International Review of the Red Cross*, September-October 1996, No. 314, p. 551-561.

State-level agencies

Parent and Child's Guide to Juvenile Records. Texas Youth Commission. (Last updated September 19, 2001). Retrieved online June 1, 2003 from: http://www.tyc.state.tx.us/programs/parentguide_records.html

"In Texas there now exists a records system that is designed to limit access to your juvenile records after you reach 21 years of age if you do not commit criminal offenses after becoming 17 years of age. The system is called 'Automatic Restriction of Access to Records.' This is in addition to your opportunity to have your records sealed and destroyed under other provisions of the Texas Family Code."

The Voluntary Adoption Registry System. Texas Department of Health-Bureau of Vital Statistics. (Not dated). Retrieved online June 1, 2003 from: <http://www.tdh.state.tx.us/bvs/car/open>

The Voluntary Adoption Registry system becomes open to adoptees, birth parents, and biological siblings once they are 18 years or older.

U. S. Federal agencies

Research Presidential Materials. National Archives and Records Administration (NARA). (Not dated). Retrieved online June 19, 2003 from: http://www.archives.gov/research_room/getting_started/research_presidential_materials.html#records

"In 1978, Congress passed the Presidential Records Act (PRA), which changed the legal status of Presidential and Vice Presidential materials. Under the PRA, the official records of the President and his staff are owned by the United States, not by the President. The Archivist is required to take custody of these records when the President leaves office, and to maintain them in a Federal depository. These records are eligible for access under the Freedom of Information Act (FOIA) five years after the President leaves office. The President may restrict access to specific kinds of information for up to 12 years after he leaves office, but after that point the records are reviewed for FOIA exemptions only. This legislation took effect on January 20, 1981, and the records of the Reagan administration were the first to be administered under this law. Staff at the Reagan Library and Bush Library can provide additional information regarding access to Presidential records in their collections."

Rules of Access to the House and Senate Records. National Archives and Records Administration (NARA). (Not dated). Retrieved online June 192003 from:
http://www.archives.gov/records_of_congress/information_for_researchers/rules_of_access.html

“Although the House and Senate regularly transfer records to the National Archives and Records Administration, these remain closed to researchers for designated periods of time: 30 years for most House records, with investigative records and records involving personal privacy closed for 50 years; 20 years for most Senate records, with a similar 50-year closure period for sensitive Senate records. Some Senate committees have instructed the Center to open selected series of records to researchers upon receipt of the records by the National Archives and Records Administration.

The records of Congress are not subject to the provisions of the Freedom of Information Act.”

APPENDIX 2

Results of LANIC Electoral Observatory Exercise (Internet Archive evaluation)

Survey of 148 URLs from LANIC's Electoral Observatory (<http://lanic.utexas.edu/info/newsroom/elections/>) run through the Internet Archive. The sample URLs covered a total of 21 elections held in Latin America between December 1998 and June 2002.

	Percent	Number	
Not in Archive	13%	19	
Dead Links	61%	91	
Doesn't Cover Critical Period (of 129 sites)	36%	47	
 <u>IN ARCHIVE:</u>	 87%	 129	
ACCESS PROBLEMS (of 129 sites in archive)			
No Access to Content	9%	12	
Limited Access to Content*	16%	20	
Less Limited Access to Content**	36%	46	
Total Content Imperfectly Able to Access	60%	78	
 LINK LEVEL ***			
1st level links with some type of hindrance	7%	9	
2nd level links with some type of hindrance	20%	26	
3rd level links with some type of hindrance	13%	17	
4th level links with some type of hindrance	9%	11	
5th level links with some type of hindrance	1%	1	
Total link hindrances****	50%	64	
 ACTIVE SITES (of 129 sites in the archive)	%	Active	Total
1999 elections	35%	22	62
2000 elections	34%	14	41
2001 elections	44%	4	9
2002 elections	100%	17	17
Total Active Sites	44%	57	129

* Both graphic and link problems prevent any but limited access to site content

** allows for access problems with link and graphics but with most of content captured

*** of 66 sites with limited or less limited access

**** of the 66 sites with imperfect access, there are 2 sites with only graphic problems but not link problems

APPENDIX 3

Nigerian Election 2003 Web Archive Links Sites crawled from April 17- May 23, 2003

Election-Related Sites

European Union. Election Observation Mission to Nigeria 2003

<http://www.eueomnigeria.org/>

Independent National Electoral Commission of Nigeria

<http://www.inecnigeria.org/>

Nigeria First. 2003 Election

<http://www.nigeriafirst.org/elections.shtml>

United Nations Electoral Assistance Project in Nigeria

<http://www.unnigeriaelections.org/>

Political Party sites

All Progressive Grand Alliance. APGA Women

<http://www.apgawomen.org/>

All Progressive Grand Alliance Foundation

<http://www.apgafoundation.org/>

Alliance for Democracy

<http://www.afrikontakt.com/alliance/>

Alliance for Democracy (U.K.)

<http://afenifere.virtualave.net/>

Democratic Socialist Movement

<http://www.socialistnigeria.org/>

National Conscience Party

<http://www.nigeriancp.net/>

New Democrats

<http://www.ndnigeria.com/>

Nigeria. Presidency. National Orientation and Public Affairs (NOPA), Abuja - [Olusegun Obasanjo]

<http://www.nopa.net>

Nigerians for Good Governance

<http://npgg.freecyberzone.com/>

Peoples Mandate Party

<http://www.peoplesmandateparty.org/>

Presidential Candidate sites

Buhari, Muhammadu

<http://www.muhammadubuhari.com/>

Buhari 2003

<http://www.buhari2003.org/>

Buhari.org

<http://www.buhari.org/>

Buhari for President 2003

<http://www.mzuhari.com/>

Buhari - Okadigbo Campaign

<http://buhariokadigbo.com/>

Buhari-Okadigbo Campaign Organisation, UK & Europe

<http://www.buhari-okadigbo.com/>

Buhari / Okadigbo Campaign Organization - All Nigeria Peoples Party, ANPPUSA, Inc.

<http://www.anppusa.org/>

Nwachukwu, Ike Omar Sanda
<http://www.ikenwachukwu.com/>

Nwobodo, James Ifeanyichukwu
<http://www.jimnwobodo.com/>

[Nwodo] Chief John Nnia Nwodo, Jr.
<http://www.johnnwodo2003.org/>

Obasanjo, Olusegun
<http://www.olusegun-obasanjo.com/>

Okadigbo, Chuba
<http://www.okadigbo4president.com/>

Jibril, Sarah
<http://www.sarahjibril4president.org/>

Rimi, Dr. Mohammed Abubakar
<http://www.rimionline.com/>

Gubernatorial Candidate Sites

AKWA-IBOM

[Nkanga] Idongesit Okon Nkanga for Governor
<http://www.hope2003.org/>

ANAMBRA

Uzodike, Ajulu
<http://www.ajuluforanambragovernor.com/>

BENUE

Unongo, Wantaragh Paul Iyorpuu
<http://www.unongo.com/>

ENUGU

Nnamani, Dr. Chimaroke Ogbannaya
<http://www.ebeano.org/>

Aniagolu, Loretta
<http://www.aniagolu.org>

KWARA

Lawal, Mohammed
<http://www.lafoga.org/>

NASARAWA

Adamu, Governor Abdullahi
<http://www.abdullahiadamu.com/>

Daniel, Otunba Gbenga
<http://www.otunbagbengadaniel.org/>

Agagu, Dr. Olusegun
<http://www.agagu.com/>

Curatorial Monitored Nigerian Websites

The sites were first monitored by checking the front pages only and printing significant pages such as, front page, press releases, interviews and events. Starting May 7th, front pages and significant pages were checked for changes.

Political Parties

Site NAME	Site URL	4/15	4/16-17	4/23-24	4/29-30	5/7-8	5/13	Comments
APGA Women	apgawomen.org				checked	no change	no change	
All Progressives Grand Alliance (AGPA)	apgafoundation.org/				checked	no change	no change	
Alliance for Democracy	afrikontakt.com/alliance/				checked	no change	no change	
Allience for Democracy-UK	http://afenifere.virtualave.net/				checked	no change	no change	
Democratic Socialist Movement	socialistnigeria.org/				checked	changed*	no change	* front page
National Conscience Party	nigeriancp.net/		checked		no change	changed*	changed**	* front page ** no link to candidates
New Democrats	ndnigeria.com/				checked	no change	no change	
Nigeria. Presidency. National Orientation and Public Affairs (NOPA)	nopa.net	checked		changed	no change	no change	no change	
Nigerians for Good Governance	freecyberzone.com/				checked	no change	no change	
Peoples Mandate Party	peoplesmandateparty.org/				checked	no change	no change	

Out of 10 Political parties sites, 3 have changed. NOPA site changed after the elections.

Presidential Candidates

Site NAME	Site URL	4/15	4/16-17	4/23-24	4/29-30	5/7-8	5/13	Comments
Buhari, Muhammadu	muhammadubuhari.com/	checked		no change	changed*	changed*	no change	* front page
Buhari, Muhammadu	buhari2003.org	checked	changed	changed	changed	changed	changed*	* virus attached
Buhari, Muhammadu	mbuhari.com			checked	changed*	changed*	no change	* front page
Buhari, Muhammadu	buhariokadigbo.com/		checked	no change	no change	no change	no change*	*was down in the morning
Buhari, Muhammadu	buhari.org/pages/1/index.htm		checked		changed*	changed**	no change	* front page ** many changes
Jibril, Sarah	sarahjibril4president.org				checked	no change	no change	
Nwachukwu, Ike Omar Sanda	ikenwachukwu.com/				checked	no change	no change	
Nwobodo, James Ifeanyichukwu	jimnwobodo.com/		checked		no change	no change	no change	
[Nwodo] Chief John Nnia Nwodo, Jr.	johnnwodo2003.org				checked	no change	no change	
Obasanjo, Olusegun	olusegun-obasanjo.com/		checked	changed*	no change	no change	no change	*New presidential page up
Okadigbo, Chuba	okadigbo4president.com/intimation.htm	checked		no change	no change	no change	no change	
Rimi, Dr. Mohammed Abubakar	rimionline.com/				checked	no change	no change	

Out of 12 Presidential sites, 5 have changed. Every time Buhari2003.org was checked it had changed.

Gubernatorial Candidates

Site NAME	Site URL	4/15	4/16-17	4/23-24	4/29-30	5/7-8	5/13	Comments
Idongesite Nkanga	hope2003.org		checked		no change	no change	no change	
Ajulu Uzodike	ajuluforanambragovernor.com		checked		no change	no change	no change	
Wantaragh Paul Iyorpuu Unongo	unongo.com		checked		no change	no change	no change	
Osagie Obayuwana	nigeriancp.net/edo.html			checked	no change	no change	no change	
Femi Falana	nigeriancp.net/ekiti.html			checked	no change	no change	no change	
Chief Loretta Aniagolu	aniagolu.org		checked		site dead*	still dead*	site dead**	*files listed **no files listed
Dr. Chimaroke Ogbannaya Nnamani	ebeano.org		checked		changed*	no change	no change	*front page
Mohammed Lawal	lafoga.org			checked	site down	site up*	no change	*no change from 4/23
Adewunmi Abassi	nigeriancp.net/lagos.html			checked	no change	no change	no change	
Governor Abdullahi Adamu	abdullahiadamu.com			checked	changed*	no change	changed*	*many changes
Ogbeni Lanre Banjo	nigeriancp.net/ogun.html			checked	no change	no change	no change	
Otunba Gbenga Daniel	otunbagbengadaniel.org/			checked	no change	no change	changed*	*changes in the events section
Oyekan Arige	nigeriancp.net/ondo.html			checked	no change	no change	no change	
Dr. Olusegun Agagu	agagu.com/			checked	no change	no change	no change	
Oyebade Olowogboiga	nigeriancp.net/osun.html			checked	no change	no change	no change	
Femi Aborisade	nigeriancp.net/oyo.html			checked	no change	no change	no change	

Out of 16 Gubernatorial sites, 4 have changed.

APPENDIX 4

Timing Exercise

Measuring rates of change of Web sites based on typology

Using several tools, including the HTTrack Website Copier, the following 21 sites from Latin America were monitored for content changes during a two-week period (April 18 - May 2, 2003). The sample of 21 sites was designed to be as broad as possible in terms of both content (representing political views across a broad spectrum, ranging from mainstream groups to insurgencies, formal and informal groups, etc.) and form (mime types and file formats, small sites and large sites, etc.).

In terms of the typology used for this exercise, the following conclusions emerge:

- Party, candidate, and electoral coverage sites typically have regular or frequent content updates in the period leading up to the elections.
- Party sites for groups that are not engaged in a current electoral campaign are updated infrequently, with the exception of large, long-established, and well-endowed parties, like the PRI in Mexico.
- Sites that have a section containing news items, in this case including the alternative media and some of the New Social Movement sites, tend to be updated more frequently, in most cases daily or even multiple times during the day.
- New Social Movement sites tend to be more or less active in terms of content updates in relation to how current their "cause" is.

For each of the sites listed below, the number to the right of the site name is the number of days during the 10 day (M-F for two weeks) measuring period that the site contents were changed or updated.

Candidates & Electoral Coverage

All four sites in this category pertained to a "current" electoral campaign, in this case Argentina, with the election itself falling midway through the exercise period. As expected, the sites were very active: three of the four had content changes on a daily basis every single day during the exercise period. The Menem site changed on only two occasions.

- Kirchner
<http://www.kirchnerpresidente.com.ar/kirchner/> 10
- Menem
<http://www.carlosmenem.com/> 2
- La Nacion Suplemento Electoral
<http://www.lanacion.com.ar/coberturaespecial/lacarrerapresidencial/> 10
- UOL Suplemento Electoral
<http://www.uolsinectis.com.ar/especiales/elecciones/> 10

Political Parties

The four parties range across the political spectrum, and are a mix of "in power" and "opposition". Mexico's PRI party site had daily content changes. On the other extreme, Guatemala's FRG site had no content changes at all during the exercise period. The FMLN and FSLN sites are both very large and extensive, and had content changes intermittently throughout the measurement period.

- FSLN
<http://www.fsln-nicaragua.com/> 1
- FRG
<http://www.frg.org.gt/inicio.htm> 0
- FMLN
<http://www.fmln.org.sv/> 4
- PRI
<http://www.pri.org.mx/principal/PRI.htm> 10

Alternative Political Media

Both of these sites are very active; this exercise confirmed that the sites have multiple content changes and updates on a daily basis.

- Politica y Actualidad
<http://www.politicayactualidad.com/index.asp> 10
- Argentina Centro de Medios Independientes
<http://argentina.indymedia.org/> 10

Insurgencies

The Mexican FZLN insurgency appears to be quite active; during the exercise period, content updates were registered on about half of the days. The Movimiento Bolivariano site is affiliated with the Colombian FARC guerrilla force; none of the pages on this site have been updated since November 2002.

- Movimiento Bolivariano Colombia
<http://www.movimientobolivariano.org/> 0
- FZLN Mexico
<http://www.fzln.org.mx/> 5

"New Social Movements"

Nine sites were chosen to represent this broad category. Two of the sites had no content changes at all during this period; five of the sites changed four times or less during this period; and two, which had a large amount of news coverage related to their area of interest, changed daily or nearly every day.

- Antiescualidos.com
<http://www.antiescualidos.com/indexnew.html> 1
 - Asamblea Popular Revolucionaria
<http://www.mbr200.com/> 1
 - Chavistas.com
<http://www.chavistas.com/> 10
 - Red Bolivariana
<http://www.redbolivariana.com/> 7
 - NuevasBases.org
<http://www.nuevasbases.org/> 2
 - Cordoba Nexo
<http://www.cordobanexo.com.ar/> 0
 - El Corralito
<http://www.elcorralito.com/principal1.htm> 0
 - confinesociales.org
<http://www.confinesociales.org/> 4
 - Asociacion Conciencia
<http://www.concienciadigital.com.ar/> 4
-

Detailed Results

Site	Last updated as of 4/18	21-Apr	22-Apr	23-Apr	24-Apr	25-Apr	28-Apr	29-Apr	30-Apr	1-May	2-May
http://www.kirchnerpresidente.com.ar/kirchner/	17-Apr	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
http://www.carlosmenem.com/	early April	no	yes	no	no	no	no	no	no	no	Yes
http://www.lanacion.com.ar/coberturaespecial/lacarrerapresidencial/	18-Apr	yes	yes	yes	yes	yes	yes	yes	yes	yes	Yes
http://www.uolsinectis.com.ar/especiales/elecciones/	18-Apr	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
http://www.fsln-nicaragua.com/	13-Apr	no	n/a	no	no	no	yes	no	no	no	no
http://www.frg.org.gt/inicio.htm	n/a	no	no	no	no	no	no	no	no	no	no
http://www.fmln.org.sv/	8-Apr	no	no	no	yes	yes	yes	no	no	no	yes
http://www.pri.org.mx/principal/PRI.htm	n/a	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
http://www.politicayactualidad.com/index.asp	17-Apr	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
http://argentina.indymedia.org/	18-Apr	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
http://www.movimientobolivariano.org/	prior to 2003	no	no	no	n/a	n/a	no	no	no	no	no
http://www.fzln.org.mx/	14-Apr	no	no	yes	yes	yes	yes	no	yes	no	no
http://www.antiescualidos.com/indexnew.html	17-Apr	yes	no	no	no	no	no	no	no	no	no
http://www.mbr200.com/	9-Apr	no	yes	no	no	no	no	no	no	no	n/a
http://www.chavistas.com/	18-Apr	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
http://www.redbolivariana.com/	15-Apr	yes	no	yes	yes	yes	yes	no	yes	no	yes
http://www.nuevasbases.org/	early April	yes	no	no	no	no	no	yes	no	no	no
http://www.cordobanexo.com.ar/	3-Apr	no	no	no	no	no	no	no	no	no	no
http://www.elcorralito.com/principal1.htm	28-Mar	no	no	no	no	no	no	no	no	no	no
http://www.confinesociales.org/	16-Apr	no	no	yes	n/a	yes	no	yes	no	yes	no
http://www.concienciadigital.com.ar/	15-Apr	no	no	no	no	no	no	yes	yes	yes	yes



6. Politics, law and economics

[6.05 Legal systems](#)

[6.10 Human rights](#)

[6.15 Politics and government](#)

[6.20 International relations](#)

[6.25 Economics](#)

[6.30 Economic and social development](#)

[6.35 Agriculture](#)

[6.40 Industry](#)

[6.45 Civil, military and mining engineering](#)

[6.50 Manufacturing and transport engineering](#)

[6.55 Materials and products](#)

[6.60 Equipment and facilities](#)

[6.65 Services](#)

[6.70 Finance and trade](#)

[6.75 Organization and management](#)

[6.80 Personnel management](#)

[6.85 Labour](#)

All rights reserved. The copyright of this web site belongs to UNESCO and the University of London Computer Centre. The information provided by this web site may be freely used and copied for educational and other non-commercial purposes, provided that any reproduction of data is accompanied by an acknowledgement of this web site as the source. Under no circumstances may copies be sold without prior written permission from the copyright holders.

© University of London Computer Centre and UNESCO 2003

6.15 Politics and government

[Back to hierarchical index](#)

Government

Narrower Term

NT1 Government policy	<i>(National policy, Public policy)</i>
NT1 Political institutions	
NT2 Heads of state	<i>(Presidency)</i>
NT1 Public administration	
NT2 Central government	<i>(Federal government, National government)</i>
NT3 Civil service	
NT4 Civil servants	<i>(Public servants)</i>
NT3 Government departments	<i>(Ministries)</i>
NT2 Governance	
NT3 Electronic governance	<i>(E-governance, Online governance)</i>
NT2 Local government	<i>(Regional government)</i>
NT3 Municipal government	<i>(City government)</i>

Internal politics

(Domestic affairs, National politics)

Narrower Term

NT1 Electoral systems	
NT2 Elections	<i>(Voting)</i>
NT2 Womens suffrage	
NT1 Parliament	<i>(Legislature)</i>
NT2 Government control	
NT2 Ombudsman	
NT1 Political crises	
NT2 Political conflicts	
NT1 Political leadership	
NT2 Politicians	
NT3 Women in politics	
NT1 Political parties	

Political doctrines

(Political ideologies)

Narrower Term

NT1 Anarchism	<i>(Nihilism)</i>
NT1 Capitalism	
NT1 Collectivism	
NT2 Communism	
NT2 Socialism	
NT1 Colonialism	
NT2 Neocolonialism	
NT1 Conservatism	<i>(Traditionalism)</i>
NT1 Federalism	
NT1 Feudalism	
NT1 Imperialism	
NT2 Colonialism	
NT3 Neocolonialism	
NT1 Internationalism	
NT1 Liberalism	<i>(Radicalism)</i>
NT1 Marxism	
NT1 Militarism	
NT1 Nationalism	
NT1 Pacifism	<i>(Antimilitarism)</i>

NT2 Conscientious objection
 NT1 Pluralism
 NT1 Regionalism
 NT1 Separatism
 NT1 Technocracy (*Meritocracy*)
 NT1 Totalitarianism (*Authoritarianism*)
 NT2 Fascism
 NT2 Nazism
 NT1 Utopia

Political movements

Narrower Term

NT1 Civil war
 NT1 Guerilla activities (*Guerilla*)
 NT1 Liberation movements
 NT2 Womens liberation movement (*Feminism, Feminist movements*)
 NT1 Nonviolence
 NT1 Oppression (*Abuse of power*)
 NT2 Resistance to oppression
 NT1 Protest movements
 NT1 Revolutionary movements
 NT1 Revolutions
 NT1 Riots

Political science

Narrower Term

NT1 Political philosophy (*Political ethics*)
 NT1 Political power (*Executive power, Judicial power, Legislative power*)
 NT1 Political theory
 NT1 Politics (*Political development, Political life, Political reform*)

Political sociology

Narrower Term

NT1 Conflict research
 NT1 Polemology (*War studies*)
 NT1 Political behaviour (*Political attitudes, Political psychology*)
 NT2 Political corruption
 NT2 Political participation (*Public participation*)
 NT1 Political communication

Political systems

(Political regimes, Political structures)

Narrower Term

NT1 Colonial countries (*Colonies*)
 NT2 Colonization
 NT2 Decolonization
 NT1 Democracy
 NT2 Parliamentary systems
 NT1 Dictatorship
 NT1 Federation (*Confederation, Federal systems*)
 NT1 Monarchy
 NT1 Newly independent states
 NT1 Republic (*Presidential systems*)
 NT1 Self government (*Autonomous states*)
 NT1 State (*National state, Sovereign state*)
 NT1 World government (*World state*)

All rights reserved. The copyright of this web site belongs to UNESCO and the University of London Computer Centre. The information provided by this web site may be freely used and copied for educational and other non-commercial purposes, provided that any reproduction of data is accompanied by an acknowledgement of this web site as the source. Under no circumstances may copies be sold without prior written permission from the copyright holders.
© University of London Computer Centre and UNESCO 2003

APPENDIX 6

Political Communications Web Archive: Test Data Input Module for MODS Descriptive Metadata

Start Time

1 Title

2. Alternative Title

3. Name

4. Abstract

5. Capture Date Range

6.1 Subjects: Controlled Vocabulary - Geographical

 a) Region b) Sub-region c) Country

6.2 Subjects: Controlled Vocabulary - Subject

 a) Topic b) Actor c) Event

6.3 Subjects: Keywords From Site

7. Language

8. Genre

9. Access Condition

10. Active URL

11. Archive URL

12. Archive

End Time

APPENDIX 7

A South-North Perspective on Web Archiving

Discussion paper for the Meeting of the Curatorship Investigation Team of the Political Communications
Web Archiving Project,
Austin, Texas, December 16, 2002.

Peter Lor
Hannes Britz
2002-11-08
Revised 2002-12-12

Introduction

The topic of this discussion paper is intellectual property issues relating to web archiving as seen from the perspective of indigenous groups. We intend first to clarify the relevance of these two terms to the Web Political Communications Archiving Project, then explore some legal and moral issues relating to the harvesting of web sites. We use the term "harvesting" as shorthand for identifying, collecting, organising, preserving and providing public access to web sites or parts thereof.

Indigenous groups

In the context of indigenous knowledge systems, "indigenous groups" usually refers to first nations, the mainly pre-literate original inhabitants of countries subsequently occupied by colonial powers, colonists, or settlers. Examples: the Inuit and Amerindians of North America, the Aborigines of Australia, the San, Khoi and Bantu speakers of Southern Africa. We do not think that this use of "indigenous peoples" is appropriate to the present project. "Indigenous" must be taken more broadly to cover peoples at all levels of technological sophistication. The people responsible for the web sites of interest to this project may be from ethnic, linguistic or other minorities, non-dominant political groupings or movements that are to a greater or lesser degree overshadowed or repressed by dominant groups. In our context we should use the term "indigenous" simply to mean "of the country" or "based in the country" or "originating in the country" concerned.

Furthermore we should look at intellectual property issues from the perspective of (a) the groups responsible for creating the web sites, and (b) the citizens or inhabitants of the countries in which web sites are set up. (Perhaps also for which they were set up. Web sites are not always operated from servers physically located in the countries concerned.) More specifically, we need to place the web sites in the context of national heritage. This implies that we should also take into account the interests of the national heritage institutions in those countries whose task it is to preserve the national documentary heritage and make it accessible in the long term.

Intellectual property

There are many forms of intellectual property and various rights thereto, not all of which are readily protected by western copyright law. In developed countries there is little doubt that websites are subject to copyright and are protected by copyright law (Harris 1998). In some developing countries the relevant legislation may not be clear on this, but in so far as they have acceded to the international copyright conventions the intellectual property of these countries will in the developed countries receive the same level of protection as that of those countries themselves. <<Look at international conventions?>>

As a point of departure I propose to assume that, with a few exceptions, all web sites we want to harvest are subject to copyright. I can think of at least two categories of exceptions: (a) cases where the creation of the web site may be considered an illegal activity by the government of the country concerned, and (b) cases where web sites are collected in terms of legal deposit.

It is a legal principle (at least in some countries) that illegal activities do not receive the protection of the law. Under repressive regimes certain political groups or activities may be outlawed. Hence the products of these activities -- their publications including their web sites, may not receive copyright protection. <<To be looked at more closely.>> This is a legalistic loophole of which institutions under the rule of law in a democratic country will not want to take advantage.

Legal deposit may have a bearing on the harvesting of web sites. I expand briefly on legal deposit because in the United States legal deposit is associated with, and may be confused with, copyright registration. Today in most countries and under international conventions legal deposit is not a prerequisite for copyright. Legal deposit is the obligation imposed by law on publishers, printers and/or other parties to deposit one or more copies of their products in designated places of legal deposit, in most cases the national library. In many countries, legal deposit legislation has been extended to cover non-print material. In some countries legal deposit now covers both discrete and online digital media, including electronic journals and also web sites. Depending on how their legislation is framed, in these countries it may be legal for a legal depository to harvest web sites without contravening copyright. This would not apply to other institutions within the country. Since legal deposit law cannot be applied extraterritorially, it would also not apply to institutions outside the country.

In South Africa, legal deposit extends to online electronic publications, including web sites. This is also true of Namibia. Unfortunately, the national libraries of South Africa and Namibia do not currently have the resources to implement the legislation. Legally, however we may require deposit and to achieve deposit we may harvest web sites - or so we think; some aspects of this still need to be clarified. However, the use of the deposited materials remains subject to our copyright law.

Leaving out the above exceptions, we can assume that all web sites are protected by copyright. Websites, as information in a tangible format, fulfil the criteria set for information to be protected by copyright legislation. These criteria are that the information must be in a tangible medium and must be controllable (Britz 1997:124). Thus harvesting them without the permission of their owners or creators is technically illegal. Why "technically"? Web sites, it is generally assumed, are put out on the web to attract as many visitors as possible, so to harvest a web site, i.e. download it to a server in order to preserve it, does not appear to be "wrong". It is, after all, for a good purpose. This brings us to the moral dimension of web site archiving.

Moral arguments

For the purposes of this section it is assumed that it would be very difficult for the creators of the political web sites of interest to the Project to monitor the Project's harvesting activities and take legal action against those carrying these out. The likelihood of being apprehended and punished is negligible. So let us assume that the fear of retribution is not a factor in our decision making. This disposes of the amoral argument that we can go ahead because we will not get caught.

Based on the natural law position it is argued that there is a strong relationship between morality and legality. This implies that moral reasoning can be used to critically evaluate, and in some cases reform, intellectual property regimes.

On the basis of this moral position it can be assumed that there may be circumstances in which it is morally justified to do something illegal. Two conditions apply. The first one is when a law is in itself immoral - for example discriminatory and oppressive legislation (as under the nazi and apartheid regimes). This might imply a moral imperative to disobey the law. The second is where there is moral justification to disobey a law. The moral justification is normally based on the outcomes of an action. For example, it may be illegal to stop and get out of your car on a freeway, but if you do so in order to assist someone who has had an accident, it would be justifiable. (Not all jurisdictions are so reasonable, of course. In the Netherlands a householder who had apprehended a burglar and locked him up in a closet to await the arrival of the police, was arrested and charged with unlawfully depriving the burglar of his liberty.) India (in 1948) as well as Pakistan (in 1975) ignored international copyright legislation to enable affordable distribution of knowledge (textbooks) to educate their citizens (Basung, 1984).

There is also a school of thought that strongly supports the view that information is a common good and that it is morally justifiable to distribute it for free. The motto "information wants to be free" often reflects the sentiments of this school (Himanen 2001). Strong support for this moral position on the free access of information comes from Barlow (.....) in his thought-provoking article "The economics of ideas" where he argues that the digitisation of information will bring about the end of intellectual property. The implications are clear: those who digitise their intellectual property will not be able to protect it - it will be 'a sinking ship' <http://ifla.org/documents/infopol/copyright/jpbarlow.html>

What then are the moral arguments in favour of archiving web sites without asking for permission?

(1) "It is a good thing to harvest web sites and make them available to political scientists, historians and others for study and research. Web sites are here today and gone tomorrow. If they are not harvested, they will be lost for ever. We are doing this in the interest of science." Impressive. But in the interest of science too, the corpses of deceased aboriginal and native persons have been removed from their burial places, deposited in museums, and put on show in glass cases. There are limits to what may be done in the interest of science.

(2) "It is a good thing to harvest web sites and preserve them for posterity. They form part of a nation's documentary/cultural heritage. More and more of our history is recorded in media other than print. Web sites are a particularly significant non-print medium. We owe it to the citizens of the countries concerned to preserve at least a representative sample."

(3) "Web sites are part of the common heritage of humankind. Even if the citizens or institutions of the country concerned take no interest in their preservation, this should be done on their behalf in any case."

(4) "In developing countries the institutions that should do this lack the capacity to harvest and preserve web sites. If we don't do it, they will be lost for ever."

(5) "In some countries the institutions that should harvest and preserve web sites are controlled by oppressive regimes. They may be prevented from carrying out this task or may be pressurised into introducing some sort of bias (e.g. bias in respect of selection and preservation decisions.). We can ensure that a representative sample of political opinion in the country is preserved and made accessible."

(6) "Because web sites are so ephemeral, there is no time to approach the copyright owners for permission to harvest their sites. Communications with groups of this nature may be slow or erratic. They may be so preoccupied by their political struggle that they are not able to respond in time to request for permission. Their communications may be obstructed or monitored by an oppressive regime. There may be language barriers. They may refuse permission because of misunderstandings or for fear of manipulation, espionage or sabotage. We would be doing them a favour if we harvested their sites in spite of their objections. We are helping the powerless to get their message across." The question arises, why should we assist this particular group to "get their message across"? Who decides?

There is an analogy to this altruistic impulse that steps over legal boundaries in order to do good: a non-governmental organisation such as Médecins sans Frontières might enter a war-torn area without the permission of the government of that country, to assist refugees in rebel-held territory. This seems like taking a noble risk. But there is another analogy. In previous centuries cultural treasures were taken from colonies and other less powerful countries and deposited in museums and other institutions in the imperial powers. The Elgin Marbles are a well-known case. Ex post facto such looting has been justified on the basis that, if they were left where they were, the cultural treasures would have suffered grave damage or been destroyed due to the negligence of their owners. Today this argument is widely rejected, and in many cases there are demands for the repatriation of cultural treasures. Clearly in this case altruism is a less credible motive.

A moral 'good' approach

The question can then indeed be asked: what is the morally correct approach to web archiving? The 'free access' argument in favour of web archiving can be as unjust as the solely economic approach favouring the strict control and use of information. The ideal position would be to find a moral balance between the ownership of information (property proviso) and access to information (access proviso). Such a position would correctly reflect the dual purpose of intellectual property design.

On the one hand it must be borne in mind that one of the basic moral imperatives of merit based justice (the Lockian view) is that creators and owners of information products and services (including those who create and own web pages) have a right to control their work as well as to be compensated (economically or otherwise) for it. On the other hand, justice, based on needs propagates the accessibility and fair use and distribution of those information products. This dual nature of intellectual property must be maintained - also with regards to web archiving.

A moral conflict can arise in those cases where the property proviso restricts the access proviso - as has been demonstrated in this paper. This implies that a choice has to be made between either access to

information or the ownership thereof, which might implies exclusion. In those cases where access to information services a societal goal, in other words where information products are seen as a common good, the moral argument must be in favour of the access provisio.

The question arises then: does web archiving serve a societal goal? As part of our national heritage it can be seen as a common good. Heritage is not only concerned with the past, but also with the future. In a sense, future use and enjoyment provide the only justification for the preservation of heritage. As we move into the future, the circumstances in which heritage materials such as political web sites saw the light recede into the past. No longer the subject of such intensely partisan goals and activities, these materials become of more general and scholarly interest, from where they can be integrated into a more balanced, mature and nuanced understanding of the making of a nation and its place in the world. It is in that sense that archived web sites will become national and ultimately international heritage. We further suggest that such an understanding is what makes heritage a common good.

Guidelines

Pragmatically, we are going to be in the business of harvesting at least some web sites without obtaining prior permission from copyright owners. If we were not, there would be little point in our being here. So, what to do if we “have to” harvest web sites without permission. The following guidelines are proposed in case where the property provision has to be overridden:

- (1) Always ask for permission first.
- (2) If this is not possible, ask for permission ex post facto.
- (3) If permission is not granted, reconsider the continued retention of and access to the material. We should develop a set of criteria to guide us in these decisions.
- (4) Make the material accessible to the original creators (i.e. the political groups or movements that set up the web sites).
- (5) Make the material accessible to scholars and institutions in the country of origin. For them, lower barriers to access that are constituted by user charges.
- (6) Take care not to play into the hands of repressive regimes.
- (7) Take care that the interpretation of the material is not of such a nature as to reinforce First World prejudices about the countries of origin.
- (8) Ensure that the interpretation of the material and research on it is not done only by US scholars, but also by scholars from the countries of origin. Encourage them to undertake research on the material by making available bursaries for postgraduate studies, stipends and visiting scholarships.
- (9) Assist institutions in the countries of origin to build capacity (technological as well as methodological and ethical) to harvest and preserve their web material themselves in future.

<<Note: It would be interesting to look at these guidelines in the context of other, more general guidelines for ethical conduct in research, such as those of the African Studies Association (2002), which in short say:

- Do no harm
- Open an full disclosure of objectives, sources of funding, methods, and anticipated outcomes
- Informed consent and confidentiality
- Reciprocity and equity
- Deposition of data and publications

The guidelines proposed above for web archiving appear to be in line with ASA guidelines 1, 4 and 5, and partly with 2 and 3.>>

References

(Still under construction.)

African Studies Association. (2002) Guidelines of the ASA for ethical conduct in research and projects in Africa. Web page, URL: www.africanstudies.org/asa_guidelines.htm. Accessed 3 December 2002.

Barlow, J.P. 1994. The economics of ideas: a framework for rethinking patents and copyrights in the information age. WIRED (2.03, March 1994). Available at URL: <http://www.ifla.org/infopol/copyright/jpbarlow.htm> Accessed December 12, 2002.

Basung, L. T. 1984. Reprinting of foreign publications in some developing countries. Journal of Philippine librarianship. (March-September, 1984): 93-99.

Britz, J.J. 1997.

Harris, L.E. 1998. Digital property: currency of the 21st century. Toronto: McGraw-Hill Ryerson.

Himanen, P. 2001. The hacker ethic and the spirit of the information age. London: Secker & Warburg.

APPENDIX 8

Technical Challenges of Web Archiving

Leslie Myrick
November 14, 2003

As a complement to decisions concerning the selection, accession and management concerns that are the focus of Curatorial Team's Collection Policy, the primary Technical Collection Management decisions also involve capture, storage, preservation, management, and access. The requirements that inform all of these decisions are bound to be complex in the case of archiving and preserving born-digital objects; more complex still in the case of Web sites. Decisions must be made concerning harvest configuration and timing, data storage models and archive file formats; data format issues; preservation strategies - whether migration, emulation, refreshing, or some combination; data access mechanisms e.g. persistent identifiers; metadata standards and cataloguing; administrative access; user access mechanisms; and quality control. In a project such as the Political Communications Web Archive (PCWA), the Long Term Resources Management, Curatorial and Technical Teams' decision-making and construction of an architecture that will assure collection, preservation and access are intricately intertwined; we will lay out our evaluation of those technical aspects of the endeavor that can be separated out for scrutiny.

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 are invoking the OAIS reference model¹ and the Trusted Digital Repository model² as the twin pillars upon which to construct a viable system for preserving access to digital objects. An OAIS-compliant, trusted repository should be modular, scalable, and tightly bound by a flexible, extensible metadata system.³ In such a repository five subsystems or entities assure preservation of ingested digital objects for the long term, and facilitate the smooth transition from SIP to AIP to DIP, with the end of rematerializing for users the archived digital objects stored in AIPs.

Many questions arise in the act of preserving digital material culled from the Web: what exactly is being archived? what are the "significant properties" that must be preserved? To what extent is it necessary to preserve the original look and feel in future access to the material?

The rhizomic nature of the Web makes the definition of the boundary of a Web site problematic - the great majority of sites point outward to other sites through hyperlinking. Should a repository turn off external links or leave them active? Archive external links or ignore them altogether? A repository's curators must distinguish those "significant properties" of a digital object that must be preserved; this could include or exclude external hyperlinks, "near files" such as stylesheets or icons that might not live on the same domain as the site, downloadable files in various formats, client-side scripting, dynamic functionality, etc.

A single Web page is a relatively discrete object with links to other HTML pages internal and/or external to the domain, along with inline or embedded video, sound files, images, graphics, ad services, stylesheets, javascripts, and perhaps database-generated material or other types of dynamic or deep Web content. And underlying the page itself is a substratum of code, whether HTML, XHTML or XML, or perhaps containing (and dependent upon) javascript, with accompanying stylesheets; or .cgi, .php, .jsp, .asp scripts or servlets, in the case of deep Web gateways. A Web archiving project should, wherever possible, archive all of the code and scripting that underlies the page. However, harvesting dynamic scripting may

¹ See the standard references e.g. the CCSDS Blue Book document *Reference Model for an Open Archival Information System*, 2002, <http://www.classic.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

² See *Trusted Digital Repositories: Attributes and Responsibilities*, RLG and OCLC, 2002. http://www.rlg.org/longterm/pm_framework.pdf

be a deep-Web pursuit that is, in most cases impossible. Most websters who use the robots.txt exclusion will undoubtedly protect the directories containing scripting from collection. This is one particular case where negotiation is necessary to preserve the site entire.

Another question Web archivists face at the outset is that of versioning. Because of the ephemeral nature of the Web and born-digital media, two sorts of versioning problems emerge in archiving: the capture and management of different versions of the Web page as modified by the creator or by a database, and the eventual refreshing of bits or the migration of archived pages by the archiving repository into different formats, either following a stated normalization criterion or as old formats become obsolete.

Undoubtedly many static pages remain static for their lifetime, but the percentage of pages modified daily or weekly, e.g. dynamic pages generated from databases or RSS feeds, is significant. Online newspapers are updated at least once daily, with the BBC perhaps at one extreme, claiming (if only because the homepage contains a dynamic datetime function) that the site is updated every minute of every day. According to a much-cited early study (2000) by Molina and Cho,⁴ who crawled a set of 720,000 pages on a daily basis over four months, 40% of all Web pages changed weekly, and 23% of .com pages changed daily. Dennis Fetterly et al crawled 330 million URLs nine times over three months and found that half the sites changed weekly.⁵ The greatest rates of change were found in pages that contained banner ads, counters and date scripts, news and stock ticker applets, or Weblogs. Jay Sethuraman et al found that 23% of their sample overall changed daily; while 40% of commercial pages changed daily. They set the half life for a Web page at ten days.⁶

Political Web sites, especially those belonging to radical groups and NGOs, are subject to spurts of activity around political events such as elections, coups-d'etat, legislation debates, and so on. Many of the URLs that will be monitored and archived by the CRL project are news portals and will thus undergo daily changes. In a similar vein, the online production of some radical NGOs might replicate the ephemeral nature of street pamphlets or graffiti. A particularly intriguing event that is surely be a target of an archiving project such as this would be the hijacking of a political Website. A typical eight-week-long snapshot crawl made by Alexa for the Internet Archive will happen upon such ephemeral occurrences only through serendipity. Is an Alexa crawl in and of itself sufficient for a selective, particularly volatile archiving project such as the Political Communications archive? Is it sufficient when supplemented by focused crawls as provided by the Internet Archive crawler, or HTTrack or wget run manually?

⁴ J. Cho and H. Garcia-Molina. The evolution of the Web and implications for an incremental crawler. In *Proc. of the 26th International Conference on Very Large Databases*, Sep. 2000.

⁵ Dennis Fetterly et al, A Large-Scale Study of the Evolution of Web Pages
<http://research.microsoft.com/aboutmsr/labs/siliconvalley/pubs/p97-fetterly/p97-fetterly.html>

⁶ Jay Sethuraman et al., Optimal Crawling Strategies for Web Search Engines
<http://www2002.org/presentations/sethuraman.pdf>

APPENDIX 9

Digital Preservation Considerations for Web Archiving

Nancy McGovern

December 2003

The domain of digital preservation has matured to the stage where there are prevailing, if not universal, practices. At full maturity, the ideal would be to manage all digital objects that are selected for long-term preservation as identical objects, without regard to the file formats contained in the objects. For now, preservation approaches continue to be largely format-specific. At this point, 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.

The file formats that are present on the majority of Web 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. There are, however, application-dependent formats and other types of formats that do not yet have defined preservation pathways, and there will always be new formats for which preservation approaches must be identified. For this project, we reviewed prevailing practice, and considered the implications for Web archiving.

Appendix 34 provides detailed MIME results from a review of Web crawls for the test sets of Web sites used for the Political Communications Web Archiving project, specifically Asian and Nigerian, with comparisons to results for other test sets. All crawls showed the same top four mime types—text/html, image/jpeg, image/gif and application/pdf.—in the same order. Those four types represented 92.7%, 99.2%, 97.8% 97.6% of all mimes for the ARL, CURL, Asia and Nigeria crawls respectively. Nigerian sites showed an even smaller percentage of text/html objects, with over half the total mime objects being jpegs or gifs, by far the highest proportion of any of the crawls. (see the Mercator crawl results in Appendix 32)

Prevailing digital preservation practice

A digital archive has several possible options for accepting file formats:

1. Limit the file formats that will be accepted by the digital archive to a subset of formats for which the archive has established procedures that are affordable and/or doable.
Considerations: This is a proscriptive approach that has the advantage of contingent the preservation activities of the digital archive to manageable options, but the disadvantage of decreasing the comprehensiveness of the digital archive collections.
2. Accept all file formats that are submitted, then seek preservation solutions for those formats that do not have a defined solution at the time of acquisition, treating all formats equally to the extent possible.
Considerations: This is an inclusive approach that carries a potential risk for the digital archive organization that depositors or users will expect the digital archive to be preserving files for which the archive has yet to develop an implemented approach.
3. 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 1 preservation for all text-based formats for an unlimited time period, and level 3 bit preservation for x type of application format for the next 5 years with review at that point.
Considerations: This is an inclusive approach that, if done well, balances the capabilities of the archive and the expectations of depositors and users for accessing the files.
4. Accept all file formats submitted, 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.
Considerations: This is an inclusive approach that ensures the archive is able to preserve the files, but may entail loss of functionality that is not acceptable for the depositors or users of some collections. This may be especially problematic if the archive is not very explicit about what it will and will not preserve.

Whichever option works best for a particular digital archive, the organization that operates the digital archive should clearly and explicitly document the selected option(s) in its preservation policy, and make its policies and procedures for the digital archive widely available to depositors and users.

Implications for Web archiving

These preservation options have implications for Web archiving projects. Web crawls, the primary means by which a Web archiving project acquires materials, when successful, take in all of the formats that are present on a target Web site. After considering the options, the ideal for the Political Communications Web Archiving project would be to accept and preserve all formats as captured. The group determined that retaining the look and feel of the Web sites is a core objective of the project.

In practice that removes the first and fourth options described above. The ideal practice for political materials would not limit the intake of formats by type (option 1) or convert acquired files to more preservable formats (option 4).

Option 2 may be more problematic for a Web archive because the interaction with users (and depositors, when appropriate) will generally be through asynchronous global access using an interface to the archive. Therefore, it is more important for the preservation activities of the archive to be very clearly-defined and unambiguous to avoid unrealistic expectations.

Option 3 is a good match for Political Communications, and likely for other Web archiving projects. There are good examples of this approach that are already in place, e.g., the levels defined for the Sunsite at Berkeley (<http://sunsite.berkeley.edu/Admin/collection.html>), the Safekept approach of PADI (<http://www.nla.gov.au/padi/safekeeping/safekeeping.html>), and Harvard's Digital Repository Services (<http://hul.harvard.edu/ois/systems/drs/policyguide.html#preservation>). There is also ICPSR's Extent of Processing Approach (http://www.icpsr.umich.edu/help/abstract.html#EXTENT_PROCESS) that could be adapted.

In practice, this approach would:

- accept all formats that were captured by crawls
- categorize the level of preservation based upon the file format type, e.g., level 1: full preservation for text, HTML, XML, etc.; level 2: file migration and reduction of loss for GIF, JPG, PDF, etc.; and level 3: "as is" retention and monitoring for application files, proprietary formats, software-dependent files, etc.
- retain access to level 2 and 3 formats while actively seeking preservation solutions that would retain the look and feel of the original files
- document the level of preservation and update the preservation status over time for users of the digital archive

As the MIME results indicate, text and image files predominate, but there are still significant numbers of application, non-standard image, and atypical text files that might present preservation problems. The former is a boon for preservation, the latter the bane - and potentially very costly. We propose establishing a matrix of file formats using the MIME content types and subtypes to document the current preservation approach for each format, assign each format type to a preservation category, and define the level of preservation support by the digital archive associated with each category.

APPENDIX 10

Risk Management for Web Resources

Nancy McGovern

December 2003

A full risk management approach to the long-term preservation of Web resources requires a complex combination of organizational and technological quantitative and qualitative measures. Risk management protocols and techniques are well developed in many domains, wherever valued assets may be threatened by natural and man-made consequences, yet preservationists were fairly slow to engage in parallel developments.

Recently, risk management has become a trend in preservation, particularly for Web resources. There are two distinct varieties of risk concerns regarding the Web. The first defines risk based upon the potential liability of an institution based upon the content of its Web site, or a Web site for which it is responsible. The second defines risk based upon the 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. In this project, we are interested in the second classification of risk.

Similarly, Web archiving may refer to two distinct types of activities: one, monitoring and capture pertaining to Web-based publications, and two, capturing entire or portions of Web sites as discrete pages contained within a boundary defined by all or a segment of a URL (e.g., all of the pages at or beneath a specified directory level). This project is interested in both types of capture, though primarily the latter. The former can be viewed as a specialized part of the latter. It should be noted that risks to the individual publications may be easier to detect and prevent than to Web sites.

There are numerous ways to measure potential risks, but it is often a combination of factors that would identify real risks. The possible combinations that indicate risk to Web resources are not finite, but change as technology, institutions, and resources change. Perceived risk is also based upon an institution's determination of acceptable loss. Documented changes in the number or size of pages, structure, or format of Web pages and sites of interest may or may not 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 for the development of risk categories and appropriate responsiveness. Risk can be measured in a number ways and at a number of levels, as discussed below.

Quantitative

Web crawlers and other tools can easily determine the actual size of Web pages and sites, and note incremental and sudden changes over time, but cannot determine the cause, nature, or impact of changes absent established rules or scales for interpreting the results. These kinds of quantitative measures may be the easiest to obtain, but are only reliable indicators in conjunction with other data or as analyzed within protocols and formulae that are appropriate to the level at which the change occurred. See [Appendix 34](#) for detailed page number and site size comparative results for political Web sites.

Page-level

A page should be evaluated as a standalone entity for characteristics that might suggest good management or risk indicators, as well as within the context of a Web site. Was it created using current and prevailing markup languages, and metadata tags? Is it well-formed? Are there identifiable and meaningful dates associated with the page in HTTP headers and/or within the page headers, tags, or textual content? Are the MIME types for the page commonly used, open source, non-proprietary? Does the page contain any known potential weak spots for remote attacks based upon CERN reports and other security monitoring sources? Answers to these questions might provide indicators that pages are well-managed and likely to be maintained, or suggest that the pages are at risk.

Link-level

Link checkers and other tools will determine the status of internal and external links, and monitor changes in the status of links over time. Sudden missing links may be traceable to badly managed Web site redesigns or upgrades, either on the site through internal links or a linked site through external links, but

other kinds of changes and fundamental changes in the content that is linked to be the target site may be more difficult to detect.

Site-level

Web site crawlers, analyzers, and site mapping tools document the structure and physical content of a site (i.e., pages, attributes within pages), and track changes to the site over time. It is harder to detect the basis for those changes and to determine if the changes equate to risk because often these changes stem from organizational or technological change that acts upon the site. An understanding of events and other change drivers is essential for capturing sites of interest. This is particularly true for political Web sites, which have a higher incidence of event-based and topical Web sites. The Nigerian election Web sites provided a good example for study. See [Appendices 32, 33, and 34](#) for results of crawls.

Server-level

Server-level changes can be harder to detect. For example, it is possible to identify the type of Web server software that a site uses, but, depending on the settings a Web master chooses, the specific version and other attributes of the software installation may be hidden from crawlers. Other significant pieces of information about the server configuration and its operation might also be hidden. From a system security perspective, these shields reflect reasonable measures to protect the site; from a risk management perspective that uses remote monitoring techniques, these protective measures may remove key indicators from monitoring, placing more emphasis on other characteristics that are more readily monitored.

Organization-level

Significant changes at the organization-level include reorganizations, major programmatic or mission changes, mergers, or dissolutions. These would occur at the administrative context and external environment layers of the Web site model presented in Figure 1. Archival organizations need to understand both to knowledgably capture Web sites. This is particularly true for political sites, many of which are event-prompted by elections, changes in government, and adverse actions by governments towards groups or individuals, etc.

Most of the technical characteristics listed above should be detectable on a local or remote Web site using available or tailored Web tools. There are organizational changes that would be significant risk triggers that are much harder or impossible to detect absent some way to retrieve and respond to events and updates. For example, a change in Web master, changes in page or site owners, and decisions about Web site management.

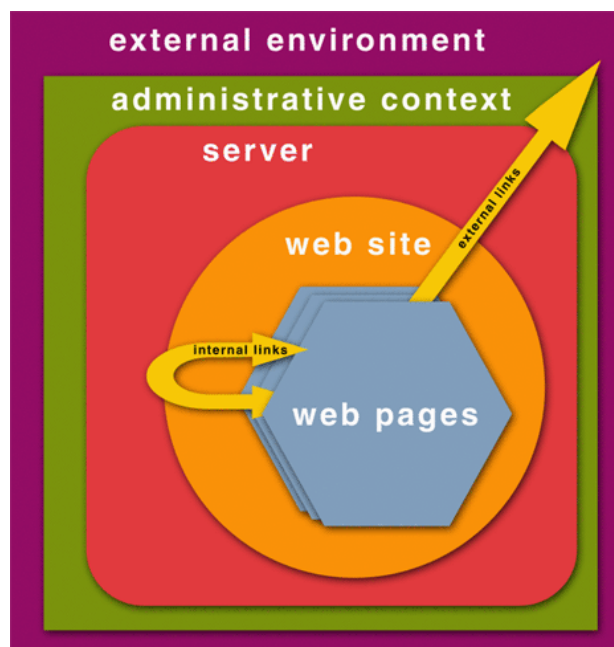


Figure 1. Virtual Remote Control Web Site Model⁷

Web site Typology

Web site typologies may be defined by content and purpose of the site, as described in the curatorial report, or by structural and other technical factors such as overall site size, number of pages, type of page, e.g. containing text-only, image-only, text plus image, dynamic indicators (forms, scripts, etc.), and incremental change over time by page and directory/location.

Using these factors in evaluating the Nigerian and Asian sites as examples of political Web sites, there are a number of characteristics that emerge (see [Appendix 34](#) for more detailed results of the evaluation):⁸

- Generally, there are fewer pages per site than for other test sites. The other sites tend to be institutional, while the political sites are often more event or subject-based sites.
- The number of pages per site, particularly the Nigerian sites, tended to remain more stable over time. Similarly, the overall size of sites is smaller and, like the number of pages, tended to remain more stable over time. These characteristics might support less frequent capture cycles for categories of political Web sites, and might make it easier to define risk parameters based upon change in the number of pages.
- There is a higher use of Apache HTTP server software for Nigerian and Asian, particularly on the Nigerian sites, than is typical across the Web, a corresponding lower use of Microsoft, and a higher use of other software. The use of less well-known software might be worrying.

Frequency of capture

Establishing the frequency of capture for individual sites is a core element of a Web archiving program. These decisions determine the size and scope of the program, and contribute significantly to the cost of archiving. Some projects have opted to take ad hoc *snapshots* of sites. A risk management approach that features preliminary site characterization and ongoing monitoring and evaluation offers the potential to schedule selective and more appropriate capture. This necessitates a combined curatorial and technological effort. As implemented in a Web archiving program, frequency of capture would be influenced the following factors:

Objectives of the organization in capturing pages/sites:

- to fully document the site by capturing all changes to the pages/sites (identified by ubiquitous monitoring against previous capture/crawl for incremental change)
- to capture significant changes to pages/sites (identified by regular monitoring against previous capture/crawl - subjective piece: what is significant?)
- to record periodic versions of the site (set uniformly for all target sites or based upon some categorization using some assessment of change over time - knowing that some content will not be captured)
- to capture a copy of pages/sites as a oneoff (because the site is short-lived, the interest in it is low, etc.)

Level of interest by archiving organization: sliding scale from essential to collection(s) to of little value to collection(s). This may also be tied to the nature of control the archiving organization might have with the creator/producer of the target site, e.g., is there an agreement with the owner to archive the site or not.

Rate of change: there are quantitative (based upon numbers and sizes - generally easy to ascertain), and qualitative (based upon techniques to gauge changes in content - generally much harder if possible at all) measures of change. Monitoring sites for a control period often identifies individual pages, clusters of pages, or directories that have higher and lower rates of change.

Scale: size of pages/site, number of pages, complexity of site, type of formats of pages. Cost may be closely tied to this (storage and backup mostly, but other organizational costs)

⁷ This is the model that Cornell devised in developing its risk management program.

⁸ The test sites that Cornell uses include ARL sites, CURL sites, a sampling of commercial and government sites for comparison purposes, and the Asia and Nigerian Election sites. The observations here are not exhaustive, but they are suggestive.

Schedules for capture should be set and kept current based upon change indicators that are weighted by these factors, including objective, level of interest, and scale.

APPENDIX 11

Web Archiving Cost Issues (Technical)

December 2003

Overview

The Technical Team was tasked with looking into the costs of Web archiving. Supplementing our review, the Curatorial report discusses staffing costs for selection and data input. The Long-Term Resource Management report incorporates cost references into their discussion of archiving activity areas. We tried to look at overall costs with a particular focus on technical cost factors.

Our review confirmed that there is a lot of interest in, but not a lot of applied work on digital preservation cost models.⁹ We opted to use a conceptual model developed by the Instruction, Research, and Information Services (IRIS) at Cornell University Library for the digital preservation management workshop (<http://www.library.cornell.edu/iris/dpworkshop/>) as a starting point, and to identify cost areas that needed more investigations within that model.

The Cornell model, which references Shelby Sanett's work, identifies three categories of cost:

Startup Costs: usually one-time expenses, including technical infrastructure (hardware, software, networks), personnel and services, and institutional overhead.

Notes for PCWA: Technical infrastructure costs for collaborative Web archiving are influenced by the choice of crawler (e.g., cost, human resources needed to operate, server capacity required to run, storage considerations of output); the spread of personnel across the collaborative (e.g., location, seniority); overhead factored based on participating members plus central unit, if appropriate.

Ongoing Costs: costs for maintaining once established (equipment, services, staffing, overheads).

Notes for PCWA: Shared costs of a collaborative may help lower these because not every member may have to sustain all of the categories.

Varying Costs: unanticipated - resulting from a major technological change, disaster of some kind, incorporation of a new preservation approach, new formats to preserve, or unexpected additions to the digital archive

Notes for PCWA: All of these could occur in a Web archiving collaborative.

Startup and ongoing costs are by definition more quantifiable. The Curatorial report contributes quantifiably to both categories; the Long-Term Resource Management report qualitatively. We identified two specific areas (with both startup and ongoing costs, plus the strong potential for varying costs) that required data gathering for our evaluation: storage and staffing. In addition, the methodology evaluation 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.

Storage Issues and Costs

A federated storage model with a primary centralized storage site supplemented by redundant mirroring of some content, and local archiving of branded content would require a range of storage systems. A high end digital library server to handle a project, such as Web archiving, that is both resource and storage-intensive would be best represented by a Sun 15K with at least of 10 Terabytes of disk storage (add Storage as necessary + tape backup); and 24-104 CPU. A Sun Center of Excellence 15K package comes with a price tag of around 4 million dollars. The Kulturaw3 server configuration is one Sun 450 for harvesting and another for storage/archiving. They use an AML/J Tape Robot for mass storage and a 1.5 Terabyte disk array as a disk cache.

⁹ For example, Brian Lavoie of the Research Department at OCLC has produced an interesting model based on roles for archiving: <http://www.oclc.org/research/projects/digipres/incentives-dp.pdf>. Shelby Sanett has framed a discussion around cost factors for preserving electronic records is adaptable to other contexts: <http://www.rlg.org/preserv/diginews/diginews7-4.html#feature2>.

Disk storage is imperative for preserving digital assets; tape should ideally only be used for emergency backup. Fortunately the trend has been for disk prices to decrease yearly by a factor of 2. Cheap disk at present can be easily negotiated at \$3,000-\$10,000 per Terabyte. Most large projects canvassed to date maintain around a 20 Terabyte target for available disk.

Technical Staffing Requirements

The technical team undertook an informal email survey of a decade of projects in production, testbed and planning stages to canvas their representatives on staffing requirements, both actual and projected (See [Appendix 12](#)). What is surprising is the patent understaffing of most projects, especially in the IT realm. The survey also makes some interesting revelations about the integration or lack thereof of librarian/archivist/curatorial/project manager personnel and IT personnel. Many projects, especially in Europe and Australia, appear to be weighted heavily in one direction or the other. This weighting has interesting repercussions on collection policy, management, metadata valuation and modes of access.

Assuming that the group managing the digital archive is not working in a vacuum, but is a component of a larger digital library structure, a minimally comfortable staffing configuration for an ambitious, selective political communications web archive would probably look something like this:

- 2 programmers/engineers initially for application development, R & D, access mechanisms; 1 programmer after the initial set up and q/a period.
- .10 system administrator; .05 network administrator; .10 DBA, .05 systems backup and recovery
- 1 Director; 1 Project Manager; 2-3 of Librarians/Curators/Archivists

Factors that will affect staffing (and costs):

- Scope of the archive: legal deposit for a national domain vs. a selective archive.
- Nature of collection/harvest: broad crawling vs. selective and/or legal deposit; push vs. pull; surface vs. deep web.
- Whether harvest/deposit is negotiated with the creators/website owners
- Extent of curatorial input in preselection vs. automated seedbed crawling.
- To what extent the material is catalogued or otherwise made available for discovery.
MARC? EAD? METS?
- Quality Control Methods
- Nature of the archival storage
Storage, maintenance, preservation: physical and logical
Refreshing/Migration/Emulation strategies
- Whether software is open-source/homegrown or uses proprietary software with service contracts.
- Access control
Nature of access interfaces.
Indexing, search and retrieval; special interfaces e.g. Wayback Machine with timeline functionality.

APPENDIX 12

Survey of Staffing Requirements/Technical Expertise in Related Projects

The following is a survey of ten web-archiving projects, roughly half using broad harvest and/or some combination of push and pull legal deposit collection, with the other half using selective harvesting. Among the selective projects four are involved in collecting and archiving political or governmental websites. Factors to consider in assessing personnel required, such as the extent of selection and cataloguing are mentioned where known.

Legal Deposit/National Libraries

PANDORA - Project Manager, 4 librarians; 1 IT expert; systems administrators; 2 FTE preservation staff are investigating long-term preservation issues.

They make a selective crawl and catalog for the most part at the website level in MARC.

<http://pandora.nla.gov.au/index.html>

Kulturarw3 - Staff is made up of 2 fulltime IT staff + 1 parttime IT staff

They undertake an automated broad crawl of the entirety of the Swedish web and do not catalog at this time. Full-text indexing is the intended means of discovery.

<http://www.kb.se/kw3/ENG/Default.htm>

WARP - employs 4 Librarians. They are in negotiations to outsource IT to a large vendor.

This is a deposit library in its early stages.

<http://warp.ndl.go.jp/>

Political Web Archives

MINERVA - Project Coordinator + 1 Librarian + 2 IT staff fulltime; part time: 2 cataloguers; 2 IT staff.

There is an expectation that more IT staff will be devoted to the project in the near future. They are in consultation with NDMSO on cataloguing and MODS issues. They contract much of the actual cataloguing out to WebArchivist.org; there is considerable consultation with the Office of General Council and Office of Strategic Initiatives.

<http://www.loc.gov/minerva/>

PRISM - Manager + 6 staff made up of 2 librarians, 2 researchers, and 2 programmers.

This is a selective archive that monitors risk factors for political websites in South East Asia and elsewhere.

<http://www.prism.cornell.edu/>

LANIC - 1 Director, 2 project and content managers, 1 programmer or IT specialist, and 5 half-time student research assistants.

This portal for Latin American political websites uses selective crawls and offers a category-based browse interface for access to web materials they curate.

<http://lanic.utexas.edu/>

Netarkivet.dk - 3 participants from the State and University Library (of Denmark), 2 participants from the digitization and web department of the Royal Library, Copenhagen; 4 participants from the Centre for Internet Research, University of Aarhus, comprising two professors and two MA student assistants.

<http://www.netarkivet.dk/index-en.htm>

Archipol - 1 Project Manager + 4 technical staff

<http://www.archipol.nl/english/project/>

Miscellaneous Archives

Wellcome Institute Medical Web Archiving Project - 1 Technical Developer; 1 Librarian for Selection, Cataloguing, deploying PANDAS.

<http://library.wellcome.ac.uk/projects/archiving.shtml>

Internet Archive - for broad crawls they outsource to Alexa, where a team of three handles the crawling: 1 Sr. Development Technician; 1 Crawl Engineer; 1 Test Engineer; for large-scale focused crawls using their own crawler a similar group would be employed. They also employ 1 Data Archivist to maintain and preserve the data and any interfaces; 1 Systems Administrator; and 1 tools developer.

<http://www.archive.org/>

APPENDIX 13

Comparative Merits of Current Methodologies

Leslie Myrick, November 14, 2003

Note: This summary does not include some results that were produced after the November 17-18 meeting at the Library of Congress, including the most recent information about the Internet Archive open source crawler and final results from the Nigerian crawls

Armed with the issues and questions presented in [Appendix 8](#), the Technical Team began its investigation of current collection methodologies and preservation programs by examining two of the larger and more successful National Deposit Library Web-archiving programs: the National Library of Australia's PANDORA project and the Kulturarw3 project at the Royal Swedish Library. (A fuller version of this evaluation can be found in [Appendix 14](#).) During our planning stage the star of the Web-archiving project of the Japanese Diet Library appeared to be rising, so we pursued interviews with them as well and include a full evaluation in [Appendix 15](#). Our Web-archiving advisor and content provider for the evaluation, the Internet Archive, especially insofar as it has partnered with the Library of Congress' MINERVA project, will be treated separately below after an initial comparison of PANDORA and Kulturarw3 in terms of storage, preservation, metadata and access issues.

Our preliminary evaluation of the national deposit libraries PANDORA project and Kulturarw3 served to sketch out two diametrically opposed approaches to collection, cataloguing, management and access; the subsequent evaluation of IA/MINERVA serves as a sort of dialectical completion insofar as the MINERVA project has consisted of a National Library (of sorts) doing both selective harvesting and culling broad swath content provided by the Internet Archive's own focused crawler and the Alexa crawler respectively.

Although the PANDORA project and Kulturarw3 are both actively involved in legal deposit collection, these Web-archiving projects' methods and practices stand at different ends of the Web-archiving spectrum on many fronts. Their differing practices reflect their respective collection and collection management policies as they relate to curation, selection, management, preservation and dissemination of Web-based materials.

While PANDORA negotiates relationships with all of its content creators and uses a selective approach to capture, supplemented by a fair amount of push technology from the creator to the archive, Kulturarw3 undertakes primarily a series of broad swath automated snapshots of all that is deemed "the Swedish Web". For crawling, PANDORA has developed in-house a harvesting/cataloguing application aptly named PANDAS that is based on Java WebObjects wrapped around the popular and free offline browser HTTrack as its harvester. The Kulturarw3 team for its part has adapted the software for the open-source Combine Harvester, originally more of a Web indexer than archival harvester, to its collecting needs.

The latter is arguably a more robust collecting application along the lines of the NEDLIB harvester, using a series of daemons or java classes to automate a number of complex harvesting tasks. On the other hand, the cataloguing modules built into the PANDAS system satisfy the particular focus of the PANDORA archive, which is to collect primarily discrete Web documents that can be catalogued in MARC and entered into their OPAC. At this point in time Kulturarw3 does not create library catalog entries for their Web material, but will depend on full-text search against the archive. This dichotomy between the librarian-centric approach to discovery and the IT-centered approach will be addressed more fully later in this report when we consider metadata and access.

Storage and Management of Archived Materials

The NLA's PANDAS implementation of HTTrack creates a mirror of the Web site and a series of logs containing a subset of HTTP header information and crawler tracking information that can be mined for descriptive and preservation metadata. Captured mirrors are stored on an HSM file system with tape backup. An original archived copy is stored, and separate work and display copies are created. Kulturarw3³ uses a multipart MIME file to hold the collection metadata, the header metadata and the file/object itself, very much like a typical output of GNU wget with headers written into the top of the file. They also store several files together in aggregates, similar to .Alexa .arc file aggregates. For storage they deploy an HSM system with 20 TB-capacity storage on DLT 7000 tape complementing disk storage of 1.5 TB.

Data Format Issues

Although the NLA has observed with interest the National Archive of Australia's practice of prescribing a limited number of file formats that it will accept into their archive, PANDORA is not planning to restrict or normalize its MIME types at this time. Because something approaching 90% of the data belongs to one of the four or five most common MIME types they plan to use preservation strategies including migration and emulation to deal with the odd 10%. More than one format is, however, kept of documents originally harvested or deposited in XML or PDF or Word in order to future-proof their continued survival.

The Swedish Library had over 400 MIME types registered in 2001; it is presently collecting close to 800 MIME types. Roughly 90% of them belong to the five most common file types. They plan to use migration and refreshing, rather than emulation, to preserve them.

Long Term Preservation Strategies

Three essential levels of preservation have been canonized in studies such as the interim report for the MINERVA prototype¹⁰ as: the preservation of bits; of content (objects) and of experience (look and feel), with a rising scale of cost and labor-investment. Similarly a triad of preservation strategies is considered by most digital archiving projects: the refreshing of bits to new media; migration to other formats or other media as they become obsolete; emulation of original soft- and hardware environments and actual soft- and hardware museums.

PANDORA is planning to wield the full range of preservation strategies: migration, emulation, hard- and software museums, or just plain refreshing for data that cannot be otherwise migrated or emulated. Kulturarw³ is depending wholly on storage and migration; they do not plan to use emulation or hoard obsolete soft- or hardware.

Metadata

Overarching management issues are the use and promulgation of standards; and where possible the adoption of open vs. proprietary standards in the operating system, software, markup, and metadata. Metadata must be standardized to allow interoperability in the case of distributed archiving systems, and it is generally good practice insofar as metadata usually ends up serving as both a management/preservation tool and an access tool. XML has become the lingua franca of not only data transfer but data and metadata management; metadata schemata such as METS, MODS, MIX and the imminent PREMIS preservation metadata schema are taking advantage of the standards-based interoperability of XML encoding.

An emerging issue in the face of descriptive and technical digital object cataloguing costs that are projected to be prohibitive is the feasibility of the automated extraction of descriptive and preservation/technical metadata from the assets themselves along with server-delivered HTTP headers that record the client/server transaction, and any additional file headers created by the harvesting module (see [Appendix 20](#)). The harvesting application should include filtering modules that can extract a good subset of metadata from the captured material itself. In the case of the Alexa/IA metadata output, the technical team wrote a series of post-processing scripts to process the .dat or metadata file that accompanies each .arc and dump the metadata into a database. It also scraped additional metadata out of the archived files - Web pages and binary files -- themselves.

A related issue in automating the population of metadata databases with programmatically extracted metadata is to what extent creator-generated metadata such as <meta> tags and <title> tags can be trusted. Before we had cracked open our first .arc file we sent some simple perl LWP crawlers after the Web sites on our seed URL lists to ascertain how many sites used <meta> DESCRIPTION and KEYWORDS tags, and how they used them. We made a similar survey of <title> tags in HTML headers. The results can be seen in [Appendices 29 and 30](#) respectively.

The misleading nature of creator-generated metadata can result from the desire to manipulate search engine ratings or from mere carelessness or error, as we show in yet another appendix, entitled the Case

¹⁰ WEB PRESERVATION PROJECT: INTERIM REPORT by William Y. Arms, January 15, 2001 <http://www.cs.cornell.edu/wya/LC-Web/interim.doc>

of the Purloined Metadata. Here, a Web page creator for a French Marxist online journal, in copying a javascript from a German sports-related Webpage, also imported the <meta> tags belonging to that page: <META CONTENT="Sport sports Baseball Basketball Beach-Volleyball Bob Boxen Bundesliga Bundesligavereine Championsleague DEL DFB DFB-Pokal Eishockey Ergebnisse Europameisterschaft Europapokal Fernsehen Football Formel1 Formel3 Fußball Golf Hallenmasters Handball Hockey Inline-Skating Leichtathletik Motorbike Motorrad Motorsport Nationalmannschaft NBA NFL NHL Reiten Rodeln Schwimmen Skifahren Skispringen Snowboard Sportarten Sportnachrichten Surfen Tennis ... [many other terms deleted here]" NAME="keywords">. (See [Appendix 31](#) for the full account).

The NLA PANDORA Project depends on MARC catalog records for discovery but has made provisions for the collection of a sizeable subset of preservation/technical metadata for long-term preservation of their Web assets. In the PANDAS system, all metadata is processed along various points of the selection/collection/preservation continuum. Collection metadata is automatically harvested from HTTP header files, but each title is described in MARC by a human cataloguer. There is adequate control over digital provenance also built into the system; e.g. any changes to the Web site made through human intervention is manually added to digital provenance metadata.

Kulturaw³ depends on the automatic generation of collection metadata from the crawler. This extracted metadata would include any information provided by the server-delivered HTTP headers, e.g. Last Modified Date, along with metadata about the capture event provided by the crawler itself. Having opted for full-text search against the pages themselves, they do not enter MARC cataloguing data into their OPAC for each title at this time.

The MINERVA/Internet Archive/WebArchivist.org Synergy

In some ways having constructed a hybrid of the harvesting methodologies of the two approaches outlined above, the MINERVA Project team, working for the most part in collaboration with the Internet Archive and WebArchivist.org, has been responsible for researching and developing tools to collect and archive born-digital objects from the Web into a series of Internet Libraries. This is an event-driven project with primarily a political focus, with discrete projects archiving Web sites that cover Election 2000, Election 2002, The 107th Congress, September 11th, the Iraqi War, and Winter Olympics 2002. For the reasons adumbrated above their project can be seen as a rich source of information about other facets of political Web archiving that were not entirely visible in the previous studies.

Harvester

The MINERVA prototype used the offline browser HTTrack run manually against twelve sites to create its testbed. For the subsequent event-based Internet Libraries they use a combination of raw Alexa crawls and focused Internet Archive crawls for their material. Alexa crawls were seen as problematic in terms of timing - many crawls had to be performed to collect an entire site. Alexa can take many days to capture a site (as we've seen from our data) and uses a breadth-first algorithm, which tends to make deep level crawling problematic. There is also no mechanism to alter crawls once they have been seeded to produce desired results.

Coverage, Scope and Size

The Election 2000 site is hosted on IA servers; 800 sites were archived daily between August 1, 2000 and January 21, 2001. It contains 800 GB of data, with 72,135,149 valid original objects. After de-duping 59,429,760 duplicate objects were removed (!); some 12.7 million valid objects remain, of which 9,972,695 are unique objects (according to checksums). User Access Mechanisms include a subject-oriented directory, where you can select sites by subjects such as Green Party Sites, Humor and Criticism Sites, along with the WayBack Machine method of choosing a URL and selecting one of many archived versions of the site.

For the September 11th Collection, they crawled daily for 3 months. The original archive contained 5 TB of data, that was honed down to 1 TB after the Internet Archive performed de-duping. It originally contained

331,299,192 objects, pared down to 55,224,374 unique objects after checksum analysis. The interface was developed by WebArchivist.org. Sites can be browsed according to four topic headings; or by a full index of sites.

Election 2002 consists of 4,000 sites archived between July 1, 2002 and November 30, 2002. The initial release consists of sites belonging to congressional and gubernatorial candidates; it will be expanded to include party and interest group sites. WebArchivist.org designed the search interface, which would be more correctly labeled a browse interface. Options now include not only browse by category but also alphabetically by state or candidate's name. The LOC would like more search capacity, and evidently WebArchivist.org is working on a searchable metadata database. An inhouse index of the archived home pages has been undertaken using Inktomi, but that does not seem to be the final solution.

Storage and File Formats

The MINERVA Internet Libraries have been for the most part hosted by the Internet Archive, with some redundancy of disk backup at the LC. The MINERVA prototype stored HTTrack mirrors of sites, but since their partnership with the Internet Archive, they have used the Alexa .arc format. An analysis of file formats collected shows that for Election 2000 roughly 92% of objects fall into the HTML/TEXT/RTF category, with 7% images and 1% PDF. For the September 11th Archive 82% of objects were HTML/TEXT/RDF, 15% were images, and 1% PDF. For Olympics 2002 85% were HTML/TEXT/RTF, 10% images and 1% PDF. A full reckoning can be found at: <http://www.archimuse.com/mw2003/papers/grotke/grotke.html>

Metadata

The Library of Congress has been one of the leaders in promulgating metadata schemes to accommodate preservation and provenance metadata, in addition to descriptive metadata for discovery, to ensure that these assets will be managed far into the future. From the outset a MARC record has been created for each Web site captured and entered into the LC OPAC. With the advent of XML-based encapsulation of metadata such as METS and MODS the LC has adapted its metadata capture to accommodate these innovations. They have contracted with WebArchivist.org to produce MODS records for the sites in the Olympics 2002, Election 2000, Election 2002 and portions of the September 11th Archive. They have also brought in the NDMSO to oversee the production of METS records for the 107th Congress Archive.

For an examination of how METS is particularly poised to accommodate the description and management of archived Web sites, see [Appendix 21](#). The skeleton for a METS object for a Web site can be found in [Appendix 22](#), and a complete METS document for a very simple Web site from the Nigerian Election crawl [*available on request*].

Internet Archive Wayback Machine Profile

The stated mission of the Internet Archive is to archive at least great portions of the entire Web using donated Alexa content ; to that end they have collected over 300 Terabytes of compressed data since their inception in 1996, adding around 12 T of data to their collection each month. A typical Alexa crawl takes around 8 weeks to complete. Data is stored on hundreds of slightly modified x86 servers running Linux. Each computer has 512Mb of memory and can hold over 1 Terabyte of data on ATA disks. The archival format is the gzipped Alexa .arc file, a 100 MB aggregate of captured files with accompanying server-delivered HTTP headers, along with a .dat file of filtered metadata from HTTP headers and from the archived files themselves; a byte-offset based index accompanies each .arc + .dat Submission Information Package. The data stream isn't altered, but the files are no longer discreet units. URLs are altered to maintain internal consistency and temporal integrity and some javascript and comments are added to the document source on-the-fly upon retrieval.

Preservation/Migration/Emulation

Maintaining copies of the Archive's collections at multiple sites (a mirror is at the modern library of Alexandria in Egypt): part of the collection is already handled this way, and we are proceeding as quickly as possible to do the same with the rest. Although DLT tape is rated to last 30 years, the industry rule of thumb is to migrate data every 10 years. Given developments in computer hardware, we will likely migrate more often than that. As advances are made in software applications, many data formats become obsolete. We will be collecting software and emulators that will aid future researchers, historians, and scholars in their research.

Metadata

"Each ARC file has a corresponding DAT file. The DAT files contain meta-information about each document; outward links that the document contains, the document file format, the document size, etc. Each host provides an index, complete.cdx, located in /0/tmp/. This index may be joined against path_index.txt, located in the same directory, for the full path of the ARC file containing the archived document. In addition to the indices located on each host, the archive also contains an archive-wide index split across 6 remote hosts. These are aliased as index1 - index6. The CDX file on each of these hosts is located in /0/wayback.cdx.gz and is formatted slightly differently than the other CDX files located on each remote host. Refer to the legend on the first line of any CDX file for information on how to interpret the data."

Access Mechanisms

Wayback Machine. "At present, the size of our Web collection is such that using it requires programming skills. However, we are hopeful about the development of tools and methods that will give the general public easy and meaningful access to our collective history. In addition to developing our own collections, we are working to promote the formation of other Internet libraries in the United States and elsewhere."

Administrative Access

Users can apply for *researcher accounts*, which give them access to the files stored files. Unix tools are made available for working with the files.

APPENDIX 14

Evaluation of Prototypes: PANDORA and Kulturarw³

Leslie Myrick

This evaluation takes as its framework the elements of the OAIS Functional Model, addressing issues of Ingest, Archival Storage, Data Management, Administration and Access as they might be applied to a system whose purpose is to harvest a preselected list of political Websites, assign metadata, manage long term preservation and provide access to a designated community through the creation of information packages.

The National Libraries of Australia and Sweden have initiated two very different systems designed to archive digital materials for legal deposit. The former uses a combination of push and pull technology and negotiates relationships with every publisher whose limited number of works have been preselected by a curatorial selector. The latter uses strictly pull processing in a broad swath harvest of what it has designated the Swedish Web without prior negotiation with the myriad of publishers whose works it collects. Both projects are dedicated to the long term preservation of national digital assets, not simply the bytes but the original look and feel of the original object.

The PANDORA Project is part of a larger digital initiative at the NLA, the Digital Services Project, initiated in 1998. They have created an architecture to manage both digitized and born digital content, comprised of five main components: the Digital Object Storage System; the Digital Archiving System; the Digital Collections Manager; a Metadata Repository and Search System; and a Persistent Identifier Resolver Service. The Kulturarw³ Project is not quite as tightly sutured into the infrastructure of the National Library of Sweden; for instance, there is no expenditure at this time for cataloguing captured materials into the library's OPAC system.

1. Selection/Harvesting Model

The National Library of Australia's PANDORA project and the National Library of Sweden's Kulturarw³ project have become archetypes of two diametrically opposed approaches to harvesting online digital publications for legal deposit. PANDORA's harvest is selective, concentrating on a predetermined list of electronic publications, while Kulturarw³ undertakes automated broad swath crawling of the Swedish Web. The NLA has a negotiated relationship with each of its publishers, obviating some of the problems associated with harvesting the deep Web, while the Swedish Library in general does not negotiate in advance, but does have contact with the publishers of Swedish online newspapers, whose sites are harvested daily or weekly and thus visited more frequently by the harvester.

Scope

PANDORA's mandate is to collect scholarly publications and publications of national interest of current and long term research value. Other material may be included as part of a cultural snapshot. In general the material should be relevant to Australia and written by an Australian author, or written by an Australian of recognized authority on a topic of international significance.

The Kulturarw³ Project has determined the boundaries of the Swedish Web through research into DNS registries. Roughly 45% of their harvest captures the .se domain, 42.5% are .com, .net, .org, .edu registered in Sweden; 12% were in the .nu domain; .05% were *suecana extreana* (externally produced sites about Sweden); 1.2% were IP addresses. Krister Persson comments: The IP addresses mentioned above might well be under the .se (or .org, .com etc.) top level domains. However the links found while collecting links out there have given back these IP addresses.

Size

In 2001, the National Library of Australia was collecting 1250 titles of Australian provenance, following selection guidelines that fit the Library's overall collection development policy. By early in 2003 they were collecting 3287 titles. At that time they had archived 5 million files in 670,000 directories, commanding 134 gigabytes of storage; in 2003 that figure was up to 400 gigabytes comprised of 14 million files. In 2001 the Library of Sweden captured some 30 million files in 1132 gigabytes of storage. The 2002 figures for Kulturarw³ show 49 million files taking up 1809 GB of disk. Their total to date is 185,95 million files in 5,571 gigabytes.

Coverage

The PANDORA project archives electronic publications with varying rates of change from monographs, with fixed content for the most part, although some evolve over time; journals whose issues appear sequentially and remain fixed as well as some whose contents change over time; and newspapers, which are sampled in snapshots; there are also some digital ephemera that do not have print equivalents, e.g. organizational and personal sites.

The Swedish Library harvests everything from the surface Web that may be reached from ordinary html <A HREF> links. Deep Web content, such as pages containing forms interfacing with database-driven Websites, is not available. What they do derive from database driven sites are static instances of dynamically created pages. In future they will also address problem pages e.g. following pdf links, and XML.

Hardware and Software

In 2001 the NLA was using a Sun E450 and two Sun E250s with a number of Sun Ultra5 workstations. Their database is Oracle. In general the PANDORA project has depended on major investments in proprietary software systems, e.g. the TeraText Content Management System and Oracle database RDBMS. However, they produced the java-based PANDAS system in house using WebObjects. The HTTrack harvester is freely available. In view of persistent clamoring at the gates for open sourcing of PANDAS, PANDORA is considering offering evaluations of PANDAS software as a preliminary to making the source available for a nominal setup/support fee.

The Swedish Library was using a Sun Solaris 450 for harvesting and a 4500 for storage/archiving, but are finding that they need to upgrade the latter. They use Sun workstations for processing and interfacing to the servers. There is no database or content management system. They use an HSM (SAM-FS) to administer the archive with an AML/J tape robot for mass storage and a 1.5 TB disk array as disk cache.

Personnel

The PANDORA project is made up of a manager and five staff members of the Electronic Texts Unit, with significant help from the IT Division, and support from the Preservation Services Branch.

Kulturarw³ employs a systems manager, two fulltime programmers with occasional input from another programmer.

2. Harvest

Some of the most pertinent ingest issues in harvesting are: 1) the determination of a harvesting model: selective or broad swath? 2) finding or building an appropriate harvester application 3) how to deal with deep Web content that cannot be automatically harvested e.g. database-driven Websites; in some cases dynamic html Web pages generated from .asp, .php, .cgi, .jsp, or .xml; or sites controlled by authentication.

Crawling involves the traversal of a site's tree of links, with parameters set to control how far along the tree to traverse, since hypothetically, an unparameterized crawl could run until it had collected the entire surface Web (should it avoid getting caught in infinite loops from various traps and black holes along the way).

Harvesting the Web is done most effectively by crawlers harnessed to databases into which metadata is extracted, having application classes, database interfaces or daemons that would control scheduling modules, the harvester's activities, link parsers, link filters, indexers and archivers.

The NEDLIB harvester, for instance, deploys interrelated daemons for each of the aforementioned functions, and a MySQL database.

<http://www.csc.fi/sovellus/nedlib/ver122/documentation122.doc>

The PANDORA project's PANDAS interface incorporates these functionalities as well as editing modules, using Oracle as its backend.

<http://pandora.nla.gov.au/manual/pandas/>

In large-scale National Library harvests of digital resources for legal deposit both push (publishers' deposit of files via FTP, Web_DAV, or portable media) and pull methods (crawling) are used in various combinations.

Crawler

In the early stages of the PANDORA project, the NLA used a version of the Harvest indexer along with WebZip, but is now deploying HTTrack.

<http://www.httrack.com/index.php>

Kulturaw³ uses a much-altered version of the Combine Harvester, an open source application written in Perl for harvesting and threshing (indexing) Web pages developed for use in the DESIRE project and further developed by Netlab for use in Nordic NWI services.

<http://www.lub.lu.se/combine/>

Crawling Frequency

For the PANDORA project crawling frequency varies according to the resource, and is decided at the point of selection of each resource. Crawls are scheduled according to the inherent dynamic of the title (infrequently for monographs; monthly, quarterly for journals), but special crawls can be initiated as needed. Because publishers sometimes push content to the library, they tend to make notifications of impending changes in format or change of publishing schedule. One-off capture is appropriate for some items (e.g. ephemera).

Kulturaw³ has done eight complete sweeps of the internet: two in 1997; three in 1998; one in 1999; one in 2000 and one in 2001. A second sweep in 2001 had to be aborted due to a complaint questioning the legality of the harvest (now resolved in their favor). They were 49 million files into their tenth sweep as of Feb 13th. Their recent move to collect newspapers means small but frequent sweeps of each of these sites, sometimes daily.

Distributed Harvest

The PANDORA project depends upon partner institutions such as the State Library of Victoria and the State Library of South Australia for some of its gathering. It is also closely associated with the Tasmanian "Our Digital Island" project.

3. Archiving and Preservation

The overarching question has to be: what exactly is being archived? And to what extent is fidelity to the original look and feel of a site important? Another preliminary question would be: what is the purpose of the archive? Mere preservation, limited access with some parts dark, or full public access? Is this a distributed archive with an interchange of SIPs and/or DIPs from partners? Or is it a monolithic enterprise that performs all the functions of ingest, storage and provision of access?

The technical issues revolve around criteria for building a trusted digital repository whose goal is preservation for the long term, including storage models; compression issues; change and version control; choice of preservation strategies such as migration, emulation, hard- and software museums and mere refreshing of data; and the assurance of the safety, integrity and authenticity of an item through mechanisms such as MD5 checksums and watermarks.

Archive File Format

PANDORA's PANDAS implementation of HTTrack creates a mirror of the Website as well as logs containing HTTP header information and crawler tracking information that can be mined for preservation metadata. Mirrors are stored on a Unix filesystem. An original archived copy is kept and separate work and display copies are created.

Kulturarw³ uses a multipart MIME file to hold the collection metadata, the header metadata and the file/object itself, very much like a typical output of GNU wget with headers written into the top of the file. They also store several files together in aggregates, much like the IA.

Data Storage Models

The PANDORA project keeps three sets of files: preservation, working and display files. The Kulturarw³ project keeps a multipart MIME file on disk and a tape backup of it.

Data Storage

PANDORA has access to disk storage expandable from 2T - 20T in the NLA's SecureData EMC Clarion FC 4700 Digital Object Storage System; they also deploy an HSM system with 8T of tape storage. Kulturarw³ uses an HSM system with 20 TB-capacity storage on DLT 7000 tape complementing disk storage of 1.5 TB.

Data Format Issues

The NLA is not planning to normalize its MIME types at this time, although the NAA does prescribe a set of allowable file formats. Because something approaching 90% of the data belongs to one of the four or five most common MIME types they plan to use preservation strategies including migration and emulation to deal with the odd 10%. More than one format is, however, kept of documents originally harvested or deposited in XML or PDF or Word in order to future-proof their continued survival.

The Swedish Library had over 400 MIME types registered in 2001; it is presently collecting close to 800 MIME types. Roughly 90% of them belong to the five most common file types. They plan to use migration and refreshing, rather than emulation, to preserve them.

Long Term Preservation Strategies

PANDORA is planning to wield the full range of preservation strategies: migration, emulation, hard- and software museums, or just plain refreshing for data that cannot be otherwise migrated or emulated.

Kulturarw³ is depending wholly on storage and migration; they do not plan to use emulation or hoard obsolete soft- or hardware.

Fidelity to the Original

One of PANDORA's driving principles is to try to retain the look and feel of the original site. One copy of each title is kept in its original format as an archival copy. Service copies are created and migrated as necessary when changes in software, file format or technology platform occur. Service copies are altered for the sake of functionality or privacy: mailtos, paypal links as well as all external links are disabled. Some unwanted parts are deleted.

Kulturarw³ shares the philosophical goal of preserving the original surfing experience. Like the Internet Archive, they have implemented temporal as well as spatial search/viewing. What remains to be developed is a full text / free text index of the indexable material. They are addressing that now as part of the NWA consortium: <http://nwa.nb.no>.

4. Management and Metadata

Overarching management issues are the use and promulgation of standards; and where possible the adoption of open vs. proprietary standards in the operating system, software, markup, and metadata.

Metadata must be standardized to allow interoperability in the case of distributed archiving systems, and it is generally good practice insofar as metadata usually ends up serving as both a management/preservation tool and an access tool.

All PANDORA metadata is processed along various points of the selection/collection/preservation continuum by PANDAS. Collection metadata is automatically harvested from HTTP header files, but each title is described in MARC by a human cataloguer. Any changes to the Website made through human intervention are manually added to digital provenance metadata.

The NLA published an extensive data dictionary as part of a larger Local Data Model in 1997. <http://pandora.nla.gov.au/dmv2.html>. A list of recommended metadata elements (1999) for the PANDORA project can be found at <http://www.nla.gov.au/preserve/pmeta.html>.

Kulturaw³ depends on the automatic generation of collection metadata from the crawler. They do not enter cataloguing data into their OPAC for each title at this time.

5. Access

Persistent Identifiers

Having made trial of both PURLs and the Handle System, in 2001 the NLA hired a consultant to reassess their needs. Her report can be found here: <http://www.nla.gov.au/initiatives/persistence/Plcontents.html>

For the time being none of DOI, PURL or the Handle System has been found adequate insofar as there is no national or international resolver service. They have therefore implemented their own library-wide PI scheme with an internal resolver service. A persistent identifier for items housed in PANDORA would take the following format: <collection id>-<work identifier>-<archive date>-<publisher's URI>-<generation code>. This scheme affords uniqueness, granularity and enough intelligence to enable grouping and relating of versions in the absence of structural metadata.

Kulturaw³ does not deploy a persistent identifier resolving system per se, but is using a 33-character filename with a timestamp as a persistent identifier internally to the system. As part of the Nordic Metadata project in 1998 they were creating unique, location-independent URNs for their archival objects. <http://www.lub.lu.se/metadata/URN-help.html>

User Access Mechanisms

For discovery through a title's associated descriptive metadata, a MARC record for each title is entered into the NLA OPAC as well as into the National Bibliographic Database. The PANDORA site includes a search engine, but better search functionality is being investigated. The search engine runs against the TeraText content management system, which replaced MetaStar.

Kulturaw³ is examining access issues. They have tested the Norwegian product FAST and will examine other software as well. They will most likely go with free text search of the entire MIME type file with metadata embedded with the html text, rather than a system that indexes and searches only metadata such as a MARC record or a finding aid.

Administrative Access

One of the principles of the PANDORA archive is immediate access to external and internal users, but this goal is balanced against the fiduciary interests of publishers when necessary. In relation to commercial publications, periods of restriction on access by external users are negotiated with publishers to protect their income from the title. Periods of restriction range from three months to five years. Access to restricted titles is provided to internal users on a single PC on which electronic copying and communication facilities have been disabled. The Library makes every effort to secure permissions over a title once it has been ingested, withdrawing it only for extenuating legal purposes e.g. because an item is banned or involved in a court case.

Four levels of access have been determined: unrestricted publications; partial commercial restriction; full commercial restriction and full restriction.

The access module for the Kulturarw³ Project is still in production. Because the National Library of Sweden is collecting Web material without negotiated contracts with publishers, their approach to rights issues will most likely resemble that of the Internet Archive, which allows for an owner of captured Web material to opt out of the archive.

6. Conclusion

The National Libraries of Australia and Sweden stand on different ends of a Web-harvesting spectrum in many respects: the former collects and fully catalogues a predetermined set of titles after negotiating with each publisher, while the latter makes a broad swath capture of the Swedish Web without previous negotiation with publishers and does not enter cataloguing information into the library OPAC. Both libraries in their capacity as holders of legal deposit are collecting and preserving a variety of materials published on the Web that are germane to their national interest.

In a recent phone conversation Margaret Phillips pointed out that the major strength of PANDORA is also its major weakness: that it is a selective archive. Because they harvest a predetermined, circumscribed list they can negotiate with every publisher; evaluate every site captured for its usefulness; and catalogue everything they harvest -- every title has a full MARC record entered into the OPAC and the NBD. 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.

The primary weakness of the selective model is that selectors are making decisions about what researchers will want in the future, basing their selection upon collection building principles for the print model, which may not be appropriate in the digital realm.

Both PANDORA and Kulturarw³ have disseminated a rich legacy of reports that can be consulted for statistics; examples of business and logical models; positions on general archival practices; as well as approaches to specific issues in ingest, management, administration, preservation and access that can be applied to archiving political communications on the Web.

7. General Sources for the PANDORA/Kulturarw3 Evaluation

National Library of Australia, Digital Archiving and Preservation at the National Library
<http://www.nla.gov.au/initiatives/digarch.html>

Margaret Phillips, Archiving the Web: The National Collection of Australian Online Publications
<http://www.ndl.go.jp/jp/information/data/nla.doc>

Pandora Business Process Model
<http://pandora.nla.gov.au/bpm.html>

Pandora Logical Data Model
<http://pandora.nla.gov.au/ldmv2.html>

Pandas Manual
<http://pandora.nla.gov.au/manual/pandas>

Margaret Phillips, Ensuring Long-term Access to Online Publications
<http://www.press.umich.edu/jep/04-04/phillips.html>

A. Arvidson et al, The Kulturarw³ Project - The Royal Swedish Web Archiw3e - An example of "complete" collection of Web pages
<http://www.ifla.org/IV/ifla66/papers/154-157e.htm>

Kulturarw³ Description: About the Project
<http://www.kb.se/kw3/ENG/Description.htm>

Kulturaw³ Statistics

<http://www.kb.se/kw3/ENG/Statistics.htm>

Kulturaw³: To Preserve the Swedish World Wide Web

<http://bibnum.bnf.fr/ecdl/2001/sweden/sld001.htm>

Additional sources of information used in this evaluation:

Telephone Conversation with Margaret Phillips, 2/3 March 2003.

Email correspondence with Krister Persson and Alan Arvidson, 3-11 March, 2003.

APPENDIX 15

Prototype Evaluation: Web Archiving Project (WARP)

Developing History:

A trial project for a three-year period that started in fiscal 2002 by the NDL (National Diet Library), Japan. The results of these projects will be used as reference to the Legal Deposit System Council.

Purpose:

To preserve information on the Internet in Japan as cultural property for the sake of future generations.

Scope:

The WARP harvests Web sites selected by the NDL. The WARP consists of two collections: the Website Collection and the Online Periodicals Collections. The Website Collection includes Web sites of governmental organizations, governmental agencies, collaborative organizations and research organizations. The definition of an online periodical for the Online Periodical Collections is a continuously published electronic resource with an identical title and consistent publication frequency.

Size:

- Online journals: 559 titles
- Governmental organizations: 6 Web sites
- Collaborative organizations: 40 Web sites

Overall procedure:

- Selecting the resources to be acquired
- Examining the structure of websites
- Negotiating and contracting for acquisition with publishers
- Specifying the unit of the information resources to be collected
- Creating preliminary metadata
- Setting harvesting, re-harvesting conditions (Including setting URL of starting page, and depth of harvesting)
- Trimming for removing the non-essential parts of the information
- Registering the individual object
- Metadata assignment.

Hardware and Software:

Server - consisting of three servers: a main server, archiving server, and web server.

These servers are also used for other electronic library services.

Main server: FUJITSU GP7000S model 650 (Enterprise 6500)

Archiving server and Web server: FUJITSU GP7000S model T1 (Netra T1)

Disk array (Storage system) - HITACHI Technology SANRISE 2800 724 GB

This storage is also used for other electronic library services.

Database software - SearchServer Version 3.7

Personnel:

Coverage:

The WARP focuses more on surface Web resources rather than deep Web resources.

Harvesting Methods:

Using a Web Robot, wget 1.5.3. (<http://www.gnu.org/software/wget/wget.html>)

Harvesting Frequency.

- Web sites - monthly
- Online Journals - depending on their publication frequency.

Archive File and Storage:

Roughly 700,000 files, about 35GB

Data Storage Models:

Data Storage:

Data Format Issues:

- HTML - 44.2%;
- JPG - 20.6%;
- GIF - 23.9%;
- PDF - 8.4%;
- others - 2.9%

Fidelity to original:

Retain original data structures of the original Web resources to display the archived resources in the same way with the original resources on the Web.

Preservation/ Migration/ Emulation:

Persistent Identifiers:

Metadata have a new URL for the Web archive and the original URL for the original site as identifiers which link to each resource.

Catalog:

Cataloguing manually before harvesting.

Metadata:

In conformity with the Dublin Core Metadata Element Set, "The NDL Metadata Element Set" was issued as the NDL standard for metadata creation in March 2001. Prior to the developments in the legal deposit system, the WARP experimentally adopts the standard and assigns metadata to collected websites and online periodicals. The metadata is filed in "Web Materials Catalogue Database"(temporary name). The metadata can be used for retrieval.

"The NDL Metadata Element Set" is based on the Dublin Core Metadata Element Set, and the NDL adopted some original qualifiers that enable mapping to the JAPAN/MARC format.

Metadata elements in the WARP are as follows: title, author, keyword, description, subject, original URL, new URL, ISSN, ISBN, NDC, language, etc.

Access Mechanisms:

The NDL plans to construct a navigation service based on metadata.

Administrative Access:

Depending on contract conditions with publishers, users can access to the contents of the WARP by Internet or by Intranet.

Look and Feel Issue:

Management Issue:

The NDL is trying to set standards such as harvesting conditions and time interval of re-harvesting by making repeated experiments. Archived resources will also be preserved in other formats such as CD-R.

The NDL manages two copies for each resource: one is for preservation, the other one for the public access.

APPENDIX 16

Harvester Evaluation

November 2003

Harvesting Specifications

Having made preliminary studies of other projects' reports on the difficulties encountered in capturing specific types of Web content such as deep Web material, frame-based, heavily javascripted, or Flash-based sites, along with an examination of reports on the efficacy of their own harvesting applications, we compiled the following list of suggested criteria/questions in evaluating crawlers in general and the three crawlers involved in two capture exercises made by our groups: harvesting selected sites pertaining to the Nigerian Election, and the LANIC time-test crawls.

- **System Configuration:** What are the system requirements? Does the harvester use a database as a backend for managing processes? Daemons? What is the API to the system?
- **Configuration/Default Settings:** What can and cannot be configured? What expertise is needed to configure it? What sort of manuals exist?
- **Problem Files:** How do they each deal with problem files e.g. dynamic pages (.asp, .php, .cgi, .jsp, DHTML, servlet-generated material, database-generated content; applets; forms): do they resolve them into html? What happens in the case of framesets, xml/xslt, downloads? Flash? Javascripts? Does the crawler accept all cookies, or can it be configured to accept only selected cookies?
- **Crawl Methods:** Do they allow for snapshots only? Incremental crawls? In incremental crawls are old files archived or overwritten? In the process of incremental replacement of old files, what are the tests for file modification (checksum, last modified date, etag)?
- **Redundancy:** Can the crawler analyze what percentage of repeated snapshots consist of redundant capture? Or ascertain the percentage of files that were not modified?
- **Archiving format:** How are Web sites and their pages archived? Do they mirror the site? Do they collect all the pages into an aggregate file?
- **Links:** Are internal links rewritten to relative links? Are external HREF links left absolute? Are paypal icons and mailtos disabled?
- **Client scripts:** Are forms, counters and active scripts disabled? In the case of javascript rollovers, how is the second image treated?
- **Metadata:** What range of metadata is captured or filtered out of the client/server transaction and/or the page itself? Where is the metadata stored? What sort of information about sites is provided in crawler logs that requires little or no post-processing? Describe each output log and report.
- **Content:** What did each crawler capture, given the expectations embedded in the configuration? How does the content collected illuminate the ephemerality factor - what changed? When? Which crawler(s) captured changes and which didn't?

Guided by these questions, and in the wake of much further study and a bit of empirical data, the following broad answers have emerged:

- A robust harvesting system such as the Internet Archive focused crawler, the Nordic Group's NEDLIB Harvester, or the Mercator Harvester harnesses a Web crawler to one or more databases and a java- or even a perl-based application that might be composed of specific classes or daemons such as a scheduler, harvester, linkparser, DNS resolver, linkfilter, metaparser and archiver to handle complex functions.
- A supplemental ad hoc crawler to provide supplemental crawls for special events as identified by curators could belong to a class of application along the lines of the NLA PANDAS system, where a simple crawler is supplemented by java-based modules for easier processing.

The ideal harvester should not only create a safe archival copy that may be as simple as an aggregate .arc file, but it should facilitate the rematerialization and access to the service version by copying or representing the original directory structure of the Web site in a zipped mirror. In order to avoid linking out to the live site, it should translate absolute links to relative links that reflect the storage path on the storage file system. It should weed out duplicate files - ideally, it should allow for incremental archiving using some combination of Etags/Last Modified server headers/checksums to prevent at least one subset of duplicate files from being harvested. It should switch off mailtos, payment devices, external links (if desired), forms.

A set of recommendations follow the three case studies below:

Case Study 1: The Alexa and IA Focused Crawlers

An Alexa .arc file is essentially an aggregation of captured HTML and associated server-delivered and crawler-generated metadata for each page bundled into a 100 MB archive file. A generic Alexa .arc is entirely promiscuous and random, according to where that particular crawler wandered in that snapshot, and how much could be packed into a 100 MB file. However, the .arc files that the CRL Project received underwent one further level of handling, to collect each of four region's URLs into its own .arc file(s). Each .arc file is complemented by a .dat (metadata) and a .cdx (index) file. The .dat file contains metadata filtered out of the http headers and the page itself, along with a list of links in each page, broken down by type. The .cdx index collects the URL, arc identifier and checksums along with other HTTP header information into a distillate that can be used to point into the .arc file. There is also a complete .cdx index for all the .arcs residing on a single Internet Archive hard disk.

How does the .arc/.dat/.cdx package stack up as an information package that could be used as an AIP in an OAIS-compliant archive? The major weakness of the .arc format that may consign it to serving as a SIP that will be transformed into an AIP and DIP is that it is not expressed as XML. On the other hand, it exists as text and binary code and is thus a simple lowest common denominator of sorts. On the negative side, the .arc format also offers no readily discernible structural metadata for a Web site, although the links originating in a single page are enumerated in the .dat file, and could be extracted to help recreate the link structure of the original site.

One immediate programming challenge lies in this fact that an .arc file does not recursively mirror the structure of a site, but is a flat aggregation of discrete pages. A possible redress is in part available on the researcher site at the IA, which offers a score of open source scripts for accessing data in an .arc file; one of them, `bin_search`, outputs a list of all the files that match a certain pattern: it could conceivably gather all the files under a root URL such as www.vaciamiento.com. Then a perl script could be written to parse out the structure into a METS structMap, assuming that an .arc contains all the files in the site; or it could be written across all .arcs belonging to the same snapshot.

Gathering adequate technical metadata for images, executables and any other binary files is another area where the .arc and .dat files will surely have to be supplemented by scripts that extract metadata from multimedia or binary headers, or by human cataloguing using other tools. For instance, total image size in bytes is available in the .dat file as well as MIME type, but the resolution and dimensions of the object are not. Where good image headers exist, a perl script could be devised to parse out information such as the bits of strings that can be found lurking in the binary data archived for each resource in the .arc file. IPTC data can be parsed out using a perl module called `Image::IPTCInfo`. Issues involved in metadata extraction from other multimedia files and executables will continue to exercise digital preservationists and will be the subject of further study.

In the positive register is the IA .dat file's helpful breakdown of parsed links into HREF links and embedded SRCs, respectively, for the two linked items; this distinction will facilitate processing of the <structMap> and <structLink> sub-elements in a METS object created to wrap Web pages' metadata and files. Two types of md5 checksum are recorded that could be used in an integrity check in a MIX <checksum> element. The IA metaparser also converts Microsoft Code Page entities found in titles into HTML character entities, although on rare occasions it strips them out instead.

Case Study 2: The NEDLIB Harvester

The Technical Team had no hands-on experience with this harvester or its output, but a number of Web-archiving projects from the Nordic realms have produced done copious documentation of the strengths and weaknesses of this system. A particularly useful description of experiences with NEDLIB can be found in the [netarkivet.dk](http://www.netarkivet.dk/rap/Webark-final-rapport-2003.pdf) final report, for instance: <http://www.netarkivet.dk/rap/Webark-final-rapport-2003.pdf>.

The NEDLIB harvester, a product of the Networked European Deposit Library Project, is primarily intended for national libraries to collect and store Web documents as part of their legal deposit activities. The software is available in the public domain, and can be downloaded from: <http://www.csc.fi/sovellus/nedlib/ver122/harvester122.tar.Z>. The components of the system are nine interrelated daemons to handle complex processes such as scheduling, link parsing, link filtering, parsing

out metadata and monitoring Webservers' performance as well as the harvesting system's performance, with a MySQL relational database backend containing nineteen tables. At this time the harvester does not include an access module/search engine, but an access interface is being developed by the Nordic Web Archive (<http://nwa.nb.no/>). The first round of capture is a full snapshot, followed by incremental updates.

Disk and software requirements include 64-bit Unix with disk; MySQL; gcc; flex; perl5; tar and gzip. The Linux filesize limit of 2G has been found to be problematic; so also MySQL's 2GB field limit. Solaris 2.6+ on the other hand can support a filesize of 1 TB and so may be preferable in cases where resource- and storage-intensive material such as video is being captured on a large scale. Minimum Processing Requirements are 450Mz Pentium III with Linux kernel 2.x; 512MB - 1 GB memory; 3 G disk space for OS; 3 G for MySQL; 5 G for harvesting activity; 2 G for accessing data in workspace.

A Good Production Configuration for NEDLIB would be 2 x 400 MHz ultraSparc II with Solaris 2.6+ ; 2 G or more memory; 1 TB tape robot for storage of archive files; 20G disk space for harvesting; 10-20G for the reqdb daemon; 30G for packing of harvested files (/tmp/); 30G for MySQL; 10G for the OS.

For a longer evaluation of NEDLIB see [Appendix 17](#).

Case Study 3. HTRACK/PANDAS

Underlying the PANDAS system is HTrack, a highly configurable offline browser that can mirror Web sites. Harvesting can be automated to a minimal degree by a scheduler in the windows GUI that allows the user to set a deferred harvesting time. It can also be set up as a cron job on Unix. There are numerous similarities between HTrack and the GNU product wget; my suspicion is that the former is wrapped around the latter with the addition of a number of stealth features such as the ability to overlook robots.txt exclusions and to spoof the user-agent type to avoid robot traps.

HTrack Features

HTrack embodies a score of features that could be considered vital in a functional archiving crawler. In a nutshell, it recursively mirrors the structure of the site; can mimic a human user and thus follow rules of proper visiting behavior are parameters involving flow control: e.g. the number of concurrent connections; the number of connections per second and the total bandwidth of the crawl; it can deploy filtering; write dynamic files to .html extensions; has proxy support; incrementally updates the archive; can save previous versions; and has copious logs full of capture metadata.

HTrack can stay on the same address, directory or domain, or move up or down from the seed URL. Mirroring depth as well as breadth of the harvest of external links are configurable, as are maximum sizes for single files, and overall Web site size. Time constraints such as the overall time limit for the crawl, the transfer rate in bytes/second can also be controlled. These parameters serve both to protect the crawler from crawling into oblivion and to disguise the robot's activity.

It can merely scan and collect information about a Web site; save only html files or only non-html files; grab html first and then non-html. In terms of building the archive, it can grab files in a list, without a recursive reconstruction of the site tree, or it can create a mirror; it can mirror Web sites in interactive mode, allowing selection of pages to harvest. Web site archives can be built in the original structure or in a user-defined structure.

For purposes of circumventing slow or faulty servers, timeout and retry rates are configurable. There are three levels of bailout: host abandon can be set to never, after a certain timeout, or when there is a traffic jam.

HTrack has well-developed link-parsing functionality: it will parse all links and test all URLs, even forbidden, if so configured. The user can choose to keep the original links or create relative or absolute links where the opposite existed. In general it is good practice to use relative links and perhaps even to concatenate them into a persistent ID, lest the live URL be activated instead of the archived one. For the sake of mirror completeness, HTrack can replace external links by error pages if external links are not to be collected, or it can simply generate 404 pages for dead internal links.

A major source of mineable metadata, the log output is deluxe and highly configurable. Choices are separate hts-ioinfo.txt and hts-log.txt files, or one integrated file, or no log at all. It also creates a directory of logs and .dat files used for updating the archive that contains an inventory list of files captured; fairly complete header information with Etags and Date Last Modified metadata, a file containing the HTTP headers and the object itself as well as MD5 checksums. All of this capture metadata is highly extractable into a metadata repository.

The major weakness of HTTrack lies in the fact that out of the box it is manually run and has only a rudimentary scheduler and parser train. There is also no database backend to manage processes or store metadata. For a longer evaluation of HTTrack see [Appendix 18](#).

Harvester Recommendations

The Long Term Resource Management Wireframe calls for federated storage of centrally collected and brokered material by a body such as the Internet Archive using focused crawling strategies and smart crawling technologies that will depend upon a robust system overseen by daemons and either with a scalable database backend or a java-based in-memory frontier for management of the crawl and storage of the metadata, supplemented by local crawling for branding of material and to assure the capture of special events. The centralized crawler would be best supplemented by a system such as PANDAS that has wrapped modular functionalities around the HTTrack crawler that could improve upon its scheduling functionality and allow for quality control and cataloguing into a metadata repository. The repository's data could be repurposed using various interfaces to output metadata packages in MARC, MODS, MIX, and METS for both discovery, preservation needs, as well as access and navigation of archived Web site or Webpage objects.

The harvester should perform an initial snapshot (or perhaps yearly snapshots?) followed by incremental, self-deduping harvests. It should archive pages that have been replaced by modified versions. For ease of service-version generation, it should be capable of writing both to an archival aggregate and to a zipped mirror of the site that can be used as a service version.

It should capture not only HTTP headers delivered with the files but should also employ a metaparser to extract or filter out metadata from the archived files themselves. Sources of this metadata could range from creator-generated <meta> tags to the <title> tag in the HTML page header to <alt> tags for links or images. It should ideally write out the extracted metadata into SQL INSERT or UPDATE statements that could be easily loaded into the metadata repository database.

It should recognize frame-based homepages and be able to render and manage frame-based sites. It should rewrite internal links to relative to the storage path of the file on the archiving filesystem, whether it is an archival file system or a service Webserver file system. It should be smart enough to rewrite often problematic creator-generated relative links to reflect the storage path. It should disable unwanted links, such as mailtos and paypal links. It should ideally be able to parse URLs out of Flash applications. It should rewrite dynamic links with file extensions such as .php and .cgi or .xml to html. It should disable forms. It should be capable of reconstructing complex client-scripted pages using the javascript document function.

A number of robust harvesters are freely available open source applications that use free databases such as MySQL or postgresSQL as their backends; thus costs would accrue not from purchase or support plans but from installation/configuration and internal support that would be included in general systems administration duties.

APPENDIX 17

Harvester Case Study: The NEDLIB Harvester

The NEDLIB harvester, a product of the Networked European Deposit Library Project, is primarily intended for use by national libraries in their collection and storage of Web documents as part of their legal deposit activities. The software is available in the public domain, and can be downloaded as a zipped tarball from:

<http://www.csc.fi/sovellus/nedlib/ver122/harvester122.tar.Z>

The manual for the release of the NEDLIB Harvester dated 21.09.2001 can be found here:

<http://www.csc.fi/sovellus/nedlib/ver122/documentation122.doc>

The components of the system consist of nine interrelated daemons and nineteen MySQL tables. At this time the harvester does not include an access module/search engine, but an access interface is being developed by the Nordic Web Archive (<http://nwa.nb.no/>); as of this writing it has been successfully paired with the FAST indexer to provide access to harvested collections.

Daemons:

The Scheduler grants new requests to harvesters by monitoring both host Web servers' performance and the internal performance of the harvesting system. The Performance Daemon computes weighted means from service times of hosts; this daemon replaces the linear queue of previous model, which had scalability and performance issues.

The Harvester asks for URLs, fetches documents into the workdir directory, or if it is initially unsuccessful, it stores them in the wqueue table. It then adds each new document to the jobqueue table for further handling by the Linkparser, Linkfilter, Metaparser and Archiver daemons. The Linkparser extracts URLs from the harvested document and adds them to the newURLs table. The Linkfilter weeds out duplicates and already harvested URLs and moves the rest into the goodURLs table. A new daemon, the rdfilter, checks all redirected URLs to prevent the robot from wandering out of the allowed domain/server/Webspace. The Doorman daemon selects an optimal set of new URLs in a text file and feeds it into the Scheduler from the request pool in the reqdb, which is managed by the Reqloader daemon.

The Metaparser uses a perl script to read documents taken from the jobqueue table and parses out metadata into metadata files that sit in the workdir.

The Archiver daemon transfers fetched documents from the workdir to the day directory, creating a separate subdirectory for each collection date. It also creates a URN from the MD5. In the most recent version there is no longer a choice between full or incremental harvest. The first round must be full, followed by incremental updates.

MySQL Tables:

The following relational tables in a MySQL database support the activities of the various daemons:

- **authority table:** storage of authentication information.
- **config table:** storage for more dynamic configuration options than allowed in the 'definitions.h' file; e.g. 'Maxdepth' field limits a depth of searching process to avoid infinite loops.
- **disallowed table:** information on disallowed URLs, from robots.txt exclusion files.
- **documents table:** information about archived documents. Used for internal purposes only.
- **domains table:** contains all allowed domain suffixes ('.fi', '.csc.fi' etc).
- **goodurls table:** these are new requests passed along by the Linkfilter.
- **hostent table:** a cache for DNS conversion from hostname to IP.

- **hosts table:** allowed and disallowed hosts for this robot.
- **internal wait table:** used for the robot's internal communication between the scheduler and the reqloader.
- **jobqueue table:** the queue for 'harvester-linkparser-metaparser-archiver' pipe, with a limit of 2000 files.
- **knownurls table:** storage for the MD5s of URLs used by the Linkfilter to test for duplicates or already fetched URLs.
- **log table:** list of failed or broken URLs with appropriate error messages.
- **newurls table:** temporary storage for all new URLs from the Linkparser (raw material for the Linkfilter).
- **rdurls table:** table of redirect URLs.
- **robohosts table:** contains the names of those hosts from which 'robot.txt' file has been collected.
- **timespace table:** information about harvesting rounds.
- **urls table:** storage for URLs of collected documents.
- **urlroots table:** seed URLs from which to start harvesting.
- **wqueue table:** This table serves as "waiting room" for URLs that should be revisited because of previous time-out failures, and so on.

The hard- and software requirements for the NEDLIB harvester have been set out on the NEDLIB homepage:

Disk and Software Requirements:

64-bit Unix with disk; MySQL; gcc; flex; perl5; tar and gzip. Linux filesize limit of 2G found to be problematic; so also MySQL's 2GB limit. Solaris 2.6+ can support 1TB filesize.

Minimum Requirments:

450Mz Pentium III with Linux kernel 2.x

512MB - 1 GB memory

3 G diskspace for OS; 3 G for MySQL; 5 G for harvesting activity; 2 G for accessing data in workspace.

A Good Production Configuration for NEDLIB:

2 x 400 MHz ultraSparc II with Solaris 2.6+

2 G or more memory

1 TB tape robot for storage of archive files

20G diskspace for harvesting

10-20G for reqdb

30G for packing of harvested files (/tmp/)

30G for MySQL

10G for OS

Memory, I/O and Storage Issues:

The NEDLIB project has come up with a rule of thumb that 1 TB of storage is needed for 30 million compressed files. Results will, of course, vary according to the nature of the files: HTML pages, Flash applications, image or video or audio files will vary immensely in storage demands.

This harvester is disk-oriented, therefore the major bottleneck will be I/O. Better throughput can be guaranteed by using fibrechannel or SCSI interfaces rather than IDE. Because memory can be an issue with MySQL, it is preferable to run it on a server separate from the harvester itself. This assumes that network can support the I/O. The NEDLIB folks suggest that at least 1G of memory be dedicated to MySQL, and warn very strongly that the default key buffer parameter **must** be increased.

Places where bottlenecks or other problems may occur have been delineated. There are some Unix limitations that will have a bearing on how NEDLIB's harvester performs, e.g. problems with the maximum number of connections -- 1024 is often the default. There is also a limit on the number of files the Unix file system can hold - where one is repeatedly archiving objects that can contain thousands of objects this is certainly a factor. They suggest an inode value of at least one million.

NEDLIB appears to fulfill a good number of the criteria posed by the Tech Team; in the absence of actually using it however, we must base any evaluations we make on other projects' results. One such project is the netarkivet.dk group, whose testbed consists of archived Web sites from the 2001 Danish Elections. They tested an early version of the NEDLIB Harvester along with the open-source crawler wget.

APPENDIX 18

Harvester Case Study: PANDAS/HTTrack

Underlying the PANDAS system is HTTrack, a highly configurable offline browser that can mirror a Web site. Harvesting can be automated to a minimal degree by a scheduler in the Windows GUI that allows the user to set a deferred harvesting time. It can also be set up as a cron job on Unix. There are numerous similarities between HTTrack and the GNU product wget; the former may very well be wrapped around the latter with the addition of a number of useful stealth features such as the ability to overlook robots.txt exclusions and to spoof the user-agent type in order to avoid robot traps.

HTTrack Features

HTTrack embodies a score of features that could be considered vital in a functional archiving crawler. In a nutshell it recursively mirrors the structure of the site; can mimic a human user; can deploy filtering; has proxy support; incrementally updates the archive; can save previous versions; and has copious logs full of capture metadata.

How it visits a site is imminently configurable: it can merely scan and collect information about a Web site using the HEAD HTTP protocol; it can save only html files or only non-html files; it can grab html first and then non-html. In terms of building the archive, it can download a Web site as a flat series of files, without a recursive reconstruction of the site tree, or it can create a mirror; mirroring can be done in interactive mode, with the harvester allowing the human user to select which pages to harvest and which to skip. Web site archives can thus be built in the original structure or in a user-defined structure.

In configuring selection parameters HTTrack can stay on the same address, directory or domain, or move up or down from the seed URL. Mirroring depth as well as breadth of the harvest of external links are configurable, as are maximum sizes for single files, and the overall size limit to be collected. Time constraints such as the overall time limit for the crawl, the transfer rate in bytes/second can also be controlled. These parameters serve both to protect the crawler from crawling into oblivion and to disguise the robot's activity. Also important in making a crawler mimic a human user and thus follow rules of proper visiting behavior are parameters involving flow control: e.g. the number of concurrent connections; the number of connections per second and the total bandwidth of the crawl.

For purposes of circumventing slow or faulty servers, timeout and retry rates are configurable. There are three levels of bailout: host abandon can be set to never, after a certain timeout, or when there is a traffic jam.

Crawling involves the transversal of links, therefore link parsing options are important. HTTrack will parse all links and test all URLs, even forbidden, if so configured. The user can choose to keep the original links or create relative or absolute links where the opposite existed. In general it is good practice to use relative links and perhaps even to concatenate them into a persistent ID, lest the live URL be activated instead of the archived one. For the sake of mirror completeness, HTTrack can replace external links by error pages if external links are not to be collected, or it can simply generate 404 pages for dead internal links.

There are a number of attractive advanced spider options: HTTrack will accept cookies or not; check the doctype if unknown; and parse java classes if desired. It will follow robots.txt and meta tags or not and can spoof a user-agent type.

The log output is deluxe and highly configurable. Choices are separate hts-iinfo.txt and hts-log.txt files, or one integrated file, or no log at all. HTTrack also creates a directory of logs and .dat files used for incremental updating, that contains lists of files; complete header information with Etags and Last Modified Date metadata for images and pages respectively, a file containing the HTTP headers and the object itself (= 2/3 of an IA .arc file, or a wget log file with embedded headers), as well as MD5 checksums. All of this capture metadata is highly minable. HTTrack can also debug HTTP headers in the logfile. The only failing is the lack of complete HTTP headers.

The major weakness of HTTrack lies in the fact that it must be run manually or by using cron scripts; it has only a rudimentary scheduler and parser train. It cannot be called a robust harvesting system insofar as there are no scheduling or parsing daemons harnessed to a database. To address some of these issues the PANDORA group at the NLA commissioned a programmer in their IT department to write a series of

java modules to wrap around HTTrack. Most of the added functionality serves the quality control, cataloguing and metadata entry needs of the librarians who work in the PANDORA project. The CRL Tech Team was been granted permission to test-drive PANDAS though release date did not occur before the project's end phase.

APPENDIX 19

Summary of Mercator crawl problems

During the investigation, several political and event-based sites were crawled, utilizing a variety of harvesters. The following is a summary of problems surfaced in using Mercator to crawl various sites. This message is drawn from informal discussion and is not a thorough evaluation.

Detailed available crawl data differed in terms of when in the crawl cycle it was collected. For example, the arl, curl and asia crawls were, in each case, the ones I analyzed for mime-type data last year. I am limited in terms of which crawls to use, because other kinds of data (e.g. page counts, etc.) are not available for all crawls, and not all the crawl data is available, because some was lost during the various Mercator crashes.

Several of Peter's crawls used are fairly late crawls (i.e. they occurred many months after crawling for those sites had started, and in the meantime, some sites had gone down, others changed content), while the Nigeria crawl data is an early crawl (i.e. from a few weeks after we had started working with the Nigerian sites). Especially for the more volatile political sites, once a crawl list has been established, basic characterization data is best taken from early crawls, because more and more of the sites become unavailable as time goes on. For example, the Nigerian crawl of May 1, 2003 has good data on 36 of the 37 sites (one site had a robots.txt exclusion). On the other hand, the 9asia crawl was done when about half a dozen of the original sites had already disappeared.

Another problem related to the above concerns the number of sites listed for the mime-type data. I eliminated from the total count any site that had a zero sized mime-counts file. However, the results may also include sites that had gone down and later came up, perhaps with completely different content (e.g. a porn site) because the domain was bought by someone else. These kinds of changes are not well-documented and it is pretty much impossible to determine them retrospectively.

There are similar problems with the page count data. Some sites are shown with an 'x' for the page count. Others show 1 or 2 pages and are most likely not the result of legitimate crawls (Mercator typically shows two pages crawled even for sites it couldn't reach). In crawl 5arl, only 107 out of 130 total sites seem to have produced meaningful page count data.

See [Appendix 32](#) for more details on the Mercator crawl.

APPENDIX 20

Feasibility of Automatic Harvest of Preservation Metadata from Crawler Log Output

Programming modules that permit the capture and filtering out of metadata from the harvesting application and from the Web site data that is collected in the harvesting process should either be built into the harvester itself or built into the system for post-processing. Arguably a full range of metadata from descriptive to technical to structural could conceivably be derived from the material collected. We concentrate here on the descriptive and administrative/technical metadata that we expect to pull out of the process, either into a metadata database repository or into some sort of complex data structure such as a series of hashes of hashes. As we will argue, the complex structure of a Web site in all its temporal versions (and for that matter of any given Web page with its parallel elements of HTML coding, javascript or other client scripting, inline images or video or audio) is best managed by an XML schema such as METS. But in cases where METS implementation is not undertaken, we recommend capturing some metadata about the sum of the parts, including total size, a list of files, the physical file structure, and the link structure of the site.

Capture/Preservation Metadata Desiderata

The technical group in consultation with the curatorial group determined that the following metadata should ideally be captured for the harvest transaction itself, along with metadata for the Web site as a whole and for each individual file:

Host server

- IP address
- Operating System/Web Server Configuration

Harvest/capture transaction

- Timestamp
- Software doing the capture
- Configuration of the software
- HTTP Response Headers:

Status, Content-Length, Content-type, Last Modified Date, Date of transaction

- Errors

Captured Files

(ALL):

- File MIME type
- Filename
- File size
- Last Modified Date
- Creating Software
- Creating Operating System/Hardware
- Checksum/Authenticity (or MM only?)

(HTML):

- Language
- Charset
- Links broken down by type
- META tags: description, keywords HTTP-EQUIV Content-Type
- Embedded scripting
- Encoding

(Multimedia + .exe + binary text downloadables): in flux

For Images: anything that can be extracted from ImageMagick (+ grep against a "strings-ed" version for ColorSpace) and marked up in a MIX extension metadata package.

For .doc, .pdf files: anything for textMD that can be extracted by rtf2txt, pdf2txt, or the Unix strings operator. Many .pdf files, for instance, contain readily harvestable embedded RDF files.

The Web site as a complex digital object

- A summary of files
- The physical file structure
- The link structure

Archive File

- Filename
- Size
- Server/Location

We undertook an evaluation to examine the metadata filtering and logging output from five applications: the IA/Alexa crawler; HTTrack; Mercator; wget and Linklint, with an eye to the range of metadata that is recorded in logs, and to how much might be extracted programmatically into a database or a complex data structure that could output a METS object for each Web site.

Internet Archive/Alexa Crawlers

An Internet Archive SIP consists of three files: the .arc itself, which contains the full text of HTML along with two sets of metadata; the .dat file, a field-value listing of metadata which has been parsed out of the HTML files, e.g. from <meta> tags, along with file offsets, i.e. the physical byte location of the captured resource within the .arc file. Additionally, for each set of arc files there is a .cdx index file with pointer information for all the .arcs in a set or on a hard drive.

The .arc File

Each .arc file contains a filedesc header similar to the following example. A key to the parsed bits follows:

```
filedesc://IA-001102.arc 0.0.0.0 19960923142103 text/plain 200 - - 0
IA-001102.arc 122
```

Key:

```
<URL><sp><IP-address><sp><Archive-date><sp><Content-type><sp><Result-
code><sp><Checksum><sp><Location><sp><Offset><sp><Filename><sp><Archive-length>
```

Next comes a long series of html files preceded by an arc header generated by the crawler and metadata taken from the http headers sent by the host. The arc header takes the following form:

```
<url><sp><ip-address><sp><archive-date><sp><content-type><sp><result-
code><sp><checksum><sp><location><sp><offset><sp><arc filename><sp><length><nl>
```

and would look like the following:

```
http://www.dryswamp.edu:80/index.html 127.10.100.2 19961104142103
text/html 200 fac069150613fe55599cc7fa88aa089d - 209 IA-001102.arc 202
```

The server-delivered http header metadata includes http version; status; date; server; content-type; last modified date; and content-length.

The actual html file follows upon the two headers:

```
<HTML>
<HEAD>HelloWorld</HEAD>
<BODY>
Hello World!!!
</BODY>
</HTML>
```

The metadata that can be extracted from this simple page is rudimentary: a) descriptive: title from the document <head>, last modified date and content type from the http headers; b) structural: none in this case, since there are no links or related files with links; c) preservation: any of the remaining bits of MD delivered by the server or generated by the crawler: IP of the server, status, content-length, identifier of the arc file, location in the arc file, date of the crawl, and so on.

The key to .dat and .cdx files

.dat and .cdx files contain the following letters:

- A canonized url
- B news group
- C rulespace category ***
- D compressed dat file offset
- F canonized frame
- G multi-column language description (* soon)
- H canonized host
- I canonized image
- J canonized jump point
- K Some weird FBIS what's changed kinda thing
- L canonized link
- M meta tags (AIF) *
- N massaged url
- P canonized path
- Q language string
- R canonized redirect
- U uniqueness ***
- V compressed arc file offset *
- X canonized url in other href tages
- Y canonized url in other src tags
- Z canonized url found in script
- a original url **
- b date **
- c old style checksum *
- d uncompressed dat file offset
- e IP **
- f frame *
- g file name
- h original host
- i image *
- j original jump point
- k new style checksum *
- l link *
- m mime type of original document *
- n arc document length *
- o port
- p original path
- r redirect *
- s response code *
- t title *
- v uncompressed arc file offset *
- x url in other href tages *
- y url in other src tags *
- z url found in script *

* in alexa-made dat file

** in alexa-made dat file meta-data line

*** future data

The .cdx file

The .cdx file contains a line which summarizes each site having the format CDX A b e a m s c k r V v D d g M n. This translates to:

```
<url><sp><date><sp><IP><sp><original URL><sp><mime type><sp><response code><sp><old style checksum><sp><new style checksum><sp><redirect><sp><compressed offset><sp><uncompressed offset><sp><compressed dat file offset><sp><uncompressed dat file offset><sp><file name><sp><meta tags><arc document length>
```

A typical .cdx entry:

```
0-0-0checkmate.com/Bugs/Insect_Habitats.html 20010424210312 209.52.183.152 0-0-0checkmate.com:80/Bugs/Insect_Habitats.html text/html 200d520038e97d7538855715ddcba613d41 30025030eeb72e9345cc2ddf8b5ff218 - 47392928145482381 4426829 15345336 DE_crawl3.20010424210104 - 635
```

The .cdx index is not an index intended for content discovery; it is essentially collects the URL, arc identifier and checksums along with other http header information into a distillate that can be used to point into the .arc file.

HTTrack

HTTrack produces 5 or 6 logs containing metadata for its own use in keeping track of what has already been captured in incremental crawling; but these can be programmatically extracted to serve as preservation metadata in a Web archive.

A. **new.lst** That merely lists the files captured in capture order. It is not particularly useful in itself, but could be used to reconstruct the physical file structure on the server with a recursive script, for instance.

B. new.txt

A tabular log showing:

```
time size/remotesize flags(update,range, filerresponse, modified, chunked, gzipped) statuscode status
(servermsg) MIME etag/date URL localfile (from URL)
<snip>
09:06:36 391/391 ---MC- 404 error ('Not%20Found')text/html
date:Sat,%2015%20Mar%202003%2014:17:21%20GMT www.vaciamiento.com/robots.txt
(from )
09:06:37 40588/40588 ---MC- 200 added ('OK') text/html
date:Sat,%2015%20Mar%202003%2014:17:21%20GMT www.vaciamiento.com/
C:/My%20Web%20Sites/vaciamiento1/www.vaciamiento.com/index.html (from )
09:06:39 811/811 ---M-- 200 added ('OK') image/gif etag:%2266c130-32b-
3d6cc2db%22 www.vaciamiento.com/images/tex-salvemosarg.gif
C:/My%20Web%20Sites/vaciamiento1/www.vaciamiento.com/images/tex-salvemosarg.gif
(from www.vaciamiento.com/)
</snip>
```

C. New.dat

A multipart MIME file containing metadata, a checksum and the content of HTML files.

N. B. HTTrack's logging module has its own ideas about whitespace - any application extracting metadata will have to parse the 200 response code out of 2003, which is apparently a concatenation of two pieces of metadata, for instance.

1) for an image file:

```
<snip>
 2 [unknown]
329a4d859b1ea195895bbe15f061ec26823 [checksum]
2003 [response code = 200, ok]
8112 [content-length 811]
OK9 [server message]
image/gif29 [ mime type]
Wed, 28 Aug 2002 12:32:27 GMT21 [last modified date]
"66c130-32b-3d6cc2db"0 [etag]
</snip>
```

2) for an HTML file:

```
<snip>
2004 [200 = response code, i.e. successful]
94402 [content-length sent 9440]
OK9 [server message]
text/html29 [file type]
Sat, 15 Mar 2003 14:22:43 GMT0 [date]
0 [unknown]
0 [unknown]
3
HTS4
9440[content-length received]
<html>
<head>
[N.B. These meta tags can be mined for additional metadaa, e.g. charset and the editor that created
the Web page]

<meta http-equiv="Content-Type" content="text/html; charset=windows-1252">
<meta name="GENERATOR" content="Microsoft FrontPage 4.0">
<meta name="ProgId" content="FrontPage.Editor.Document">
<title>Nuevo gabinete de Duhalde</title>
</head>

<body>[ . . . ]</body>
</html>
2 [unknown]
32ed4f9f49d01db15f48c1a5b19f8744703 [checksum]
</snip>
```

D. winprofile.ini on windows [= httrack.conf in unix]

An accounting of the configuration of the crawler, for instance, whether to honor robots.txt, whether to parse java files; how deep and broad to crawl; how many retries; whether to accept cookies, and so on.

This file is generated by the windows version and kept on record for subsequent crawls of the same site. On UNIX it would be manually configured by the user.

```
<snip>
Near=0
Test=0
ParseAll=1
HTMLFirst=0
Cache=1
NoRecatch=0
```

```
Dos=0
Index=0
WordIndex=0
Log=1
RemoveTimeout=0
RemoveRateout=0
FollowRobotsTxt=2
NoErrorPages=0
NoExternalPages=0
NoPwdInPages=0
NoQueryStrings=0
NoPurgeOldFiles=0
Cookies=1
CheckType=1
ParseJava=1
</snip>
```

Some of the features of the crawl under the configuration given above are that the crawler did not follow near files, did parse java, was not asked to collect the HTML before the binary resources, did not make an index, had logging turned on, accepted cookies, and so on.

E. hts-info.txt

This file logs the request and response transactions as they occur. It contains a handful of HTTP headers including server configuration, when available:

```
Server: Apache/1.3.26 (Unix) Chili!Soft-ASP/3.6.2 PHP/4.2.2 FrontPage/5.0.2.2510 mod_perl/1.27
mod_ssl/2.8.9 OpenSSL/0.9.6b
```

A fair amount can be surmised from this information: e.g. it is likely that the site uses dynamic scripting in the form of active server pages and PHP, since ChiliSoft is a significant investment.

The log that follows records the request and response transactions between the client and the server. First the crawler asks to consult the robots.txt page, then moves on to the index page:

```
<snip>
request for www.vaciamiento.com/robots.txt:
<<< GET /robots.txt HTTP/1.1
<<< Connection: close
<<< Host: www.vaciamiento.com
<<< User-Agent: Mozilla/4.05 [fr] (Win98; I)
<<< Accept: image/gif, image/x-xbitmap, image/jpeg, image/pjpeg, image/svg+xml, */*
<<< Accept-Language: en, *
<<< Accept-Charset: iso-8859-1, *
<<< Accept-Encoding: gzip, deflate, compress, identity

request for www.vaciamiento.com/:
<<< GET / HTTP/1.1
<<< Connection: close
<<< Host: www.vaciamiento.com
<<< User-Agent: Mozilla/4.05 [fr] (Win98; I)
<<< Accept: image/gif, image/x-xbitmap, image/jpeg, image/pjpeg, image/svg+xml, */*
<<< Accept-Language: en, *
<<< Accept-Charset: iso-8859-1, *
```

<<< Accept-Encoding: gzip, deflate, compress, identity

response for www.vaciamiento.com/robots.txt:

code=404

>>> HTTP/1.1 404 Not Found

>>> Date: Sat, 15 Mar 2003 14:17:21 GMT

>>> Server: Apache/1.3.26 (Unix) ChiliSoft-ASP/3.6.2 PHP/4.2.2 FrontPage/5.0.2.2510
mod_perl/1.27 mod_ssl/2.8.9 OpenSSL/0.9.6b

>>> Connection: close

>>> Transfer-Encoding: chunked

>>> Content-Type: text/html; charset=iso-8859-1

</snip>

F. new.idx

This is an index file that consists of a pointer to the file's physical location in new.dat; it is not a discovery index per se.

<snip>

www.vaciamiento.com

/corralito.htm

366659

78

//[HTML-MD5]//

C:/My Web Sites/vaciamiento1/www.vaciamiento.com/corralito.htm

371200

42

</snip>

Mercator

Mercator crawls a site or sites starting from a pool of seed URLs. As documents are downloaded, they are analyzed on the fly by a series of custom analyzer modules. The output of the modules can be piped to other analyzers or can be written to files. The following files and the included metadata are common to all the crawls being done by Cornell.

clock.000000

Information about checkpointing, in case a crawl is interrupted.

config.sx

The editable configuration file for the crawl. Among the attributes that can be changed are the Seed URLs, the Filter strings limiting the crawl, Politeness rules, and Analyzer add-ins for massaging the raw data during run-time.

Example:

```
("SeedURLs" ("http://afenifere.virtualave.net/"))
```

```
("Filter" ("Domain" ".virtualave.net"))
```

```
("AtraxMachines" ("crawler0"))
```

```
("DnsClass"
```

```
  ("mercator.dns.MercatorDNS"
```

```
    ("NameServer" "localhost")
```

```
    ("AssumeCanonicalHost" "true")
```

```
    ("CacheOracleClass"
```

```
      ("mercator.cache.ClockReplacementOracle" (("LogSize" "17"))))))))
```

```

("TimeLimitSecs" "86400")
("CheckpointFreq" "96400")
("NumberThreads" "80")
("MaxDepth" "100")
("WorkDirPath" "/misc1/cornell/kehoe/crawls/nigelec.05013/")
("StableDirPattern" ""afenifere.virtualave.net")
("FrontierClass"
("mercator.frontier.PoliteStaticFrontier"
(("QueueThreadRatio" "3.0")
("QueueClass" ("mercator.queue.BuffDiskQueue" (("PoolSize" "200"))))
("PoolBase" "pools-1051815710173")
("PolitenessFactor" "10")
("StrictPoliteness" "true")
("MinRestSecs" "0")
("MaxRestSecs" "2147483647"))))
("URLSetClass"
("mercator.urlset.ContextAwareURLSet"
(("CanonicalizeHost" "false")
("SizeLimit" "-1")
("LogURLTrace" "false")
("LogBuffSize" "21")
("LogSpineSize" "16")
("LogCacheSize" "18"))))
("Protocols"
(("ftp"
("mercator.protocol.ftp.FTPProtocol"
(("EmailAddress" "wrk1@cornell.edu")
("AllowNonASCIIinURLs" "false"))))
("http"
("mercator.protocol.http.ContextAwareMercatorProtocol"
(("UserAgentMailbox" "wrk1@cornell.edu")
("UserAgentBase" "Mercator")
("UserAgentVersion" "1.0")
("ProxyRules" ())
("SocketTimeoutSecs" "60")
("NumberTries" "3")
("AllowNonASCIIinURLs" "false")
("AcceptHeader" "null")
("ConditionalFetch" "false"))))
("Analyzers"
(("text/html"
("mercator.analyzer.html.PageData" ())
("mercator.analyzer.html.CountingLinkExtractor" ())
("mercator.analyzer.SaveDoc" ())
("mercator.analyzer.html.HttpHeaderExtractor" ()))
("image/gif" ("mercator.analyzer.gif.GifHistograms" ()))
("ftp/directory" ("mercator.analyzer.ftpdire.CountingLinkExtractor" ())))
("ExitOnEmpty" "true")
("ContinuousCrawl" "false")
("CheckpointVersion" "0")
("StartDate" "1051815709094")
("LogFiles" ("stdout" "crawl-log.txt"))
("LogRobotsCacheSize" "17")
("FilterTwice" "false")
("DocFPSetClass"

```



```

("mercator.fpset.DiskFPSet3"
 ("LogCacheSize" "-1")
 ("LogBuffSz" "18")
 ("LogBuffTblLoad" "4")
 ("BucketSizeIncr" "4"))))
("PerDocLoggerClass" ("mercator.core.FilePerDocLogger" ()))
("LogRISMemSize" "16")
("LogRISMaxSize" "20"))

```

crawl-log.txt

A log of the actions of the crawl. It also includes the config.sx file. That section is omitted from the following example:

```

Start Date: Thu May 01 12:02:33 PDT 2003
Host: crawler0-complaints-to-admin.webresearch.pa-x.dec.com
System properties:
  java.runtime.name = Java(TM) 2 Runtime Environment, Standard Edition
  java.runtime.version = 1.3.1-beta2
  java.vendor = Compaq Computer Corp.
  java.version = 1.3.1
  java.vm.info = native threads, mixed mode, 07/31/2001-09:18
  java.vm.name = Fast VM
  java.vm.vendor = Compaq Computer Corp.
  java.vm.version = 1.3.1-beta2
  os.arch = alpha
  os.name = OSF1
  os.version = V5.1

```

```

Overridden crawler attributes:
("SeedURLs" ("http://buhariokadigbo.com/"))
("Filter" ("Domain" "buhariokadigbo.com"))
[...omitted configuration...]

```

Elapsed Wait	Discovered Memory	Frontier Pages	Downloaded Pages	Unique Pages	Overall docs/s	Current docs/s	Overall KB/sec	Current KB/sec	Failures	Current Thds	DownLoad
--------------	-------------------	----------------	------------------	--------------	----------------	----------------	----------------	----------------	----------	--------------	----------

```

-----
Workers started.
: 10.1 53 29 25 24 2.47 2.47 19.6 19.6 0 78 68699
: 20.1 54 11 44 43 2.19 1.90 22.7 25.7 0 78 86148
: 30.1 54 1 54 53 1.79 1.00 18.6 10.6 0 79 92178
: 40.1 54 1 54 53 1.35 0.00 14.0 0.0 0 79 92178

```

Flushing 54 entries to /misc1/cornell/kehoe/crawls/nigelec.05013/buhariokadigbo.com/urlfpset.curr took 5 ms

```

: 50.2 54 0 56 53 1.12 0.20 11.6 2.2 0 80 92178

```

Frontier is empty -- terminating crawl

Stopping workers...

Workers stopped.

Flushing DiskFPSet to /misc1/cornell/kehoe/crawls/nigelec.05013/buhariokadigbo.com/docfpset.curr took 16 ms

contains = 54; hits1 = 0; hits2 = 1; hits3 = 0

diskLookups = 53; seeks = 0; reads = 0

Crawl terminated at Thu May 01 12:03:25 PDT 2003

Docs-null.0000.000000

This file contains the documents as received by the client, including HTTP headers and the HTML document. Multiple pages are aggregated into one or more files. The pages are delimited by a leading line in the form "-----size-in-bytes-----URL"

Example:

```
-----22127-----http://buhariokadigbo.com:80/
HTTP/1.1 200 OK
Date: Thu, 01 May 2003 18:02:35 GMT
Server: Apache/1.3.27 (Unix) mod_throttle/3.1.2 PHP/4.3.0 mod_ssl/2.8.11 OpenSSL/0.9.6g
FrontPage/5.0.2.2510
Last-Modified: Sat, 12 Apr 2003 12:59:44 GMT
ETag: "3f80b-5525-3e980dc0"
Accept-Ranges: bytes
Content-Length: 21797
Connection: close
Content-Type: text/html
```

```
<html>
```

```
<head>
```

```
<meta http-equiv="Content-Type" content="text/html; charset=windows-1252">
```

```
<meta http-equiv="Content-Language" content="en-us">
```

```
<title>Buhari-Okadigbo Home Page</title>
```

```
<meta name="GENERATOR" content="Microsoft FrontPage 4.0">
```

[SNIP]

gif-stats.log A summary of some information about the gif files encountered. This is the output of the "mercator.analyzer.gif.GifHistograms" run-time analyzer.

http-status-histo.log

A summary of HTTP status codes for all the downloaded pages.

Example:

```
===== Statistics for Checkpoint 0 =====
```

```
-9999 = download failures / no status code
```

```
-9998 = download disallowed by /robots.txt file
```

```
Total elements: 54
```

```
Aggregate stats: min = 200, mean = 200.00, max = 200
```

```
200 --> 54 (100.0%)
```

httpHeaders-0000.000000

just the HTTP headers for each page, delimited by a leading line in the form "-----size-in-bytes-----URL". The output of the "mercator.analyzer.html.HttpHeaderExtractor" run-time analyzer.

Example:

```
-----22127-----http://buhariokadigbo.com:80/
HTTP/1.1 200 OK
Date: Thu, 01 May 2003 18:02:35 GMT
Server: Apache/1.3.27 (Unix) mod_throttle/3.1.2 PHP/4.3.0 mod_ssl/2.8.11 OpenSSL/0.9.6g
FrontPage/5.0.2.2510
Last-Modified: Sat, 12 Apr 2003 12:59:44 GMT
ETag: "3f80b-5525-3e980dc0"
Accept-Ranges: bytes
Content-Length: 21797
Connection: close
Content-Type: text/html
-----6945-----http://buhariokadigbo.com:80/Main/feedback.htm
HTTP/1.1 200 OK
Date: Thu, 01 May 2003 18:02:37 GMT
Server: Apache/1.3.27 (Unix) mod_throttle/3.1.2 PHP/4.3.0 mod_ssl/2.8.11 OpenSSL/0.9.6g
FrontPage/5.0.2.2510
Last-Modified: Fri, 11 Apr 2003 01:09:47 GMT
ETag: "edba8-19d8-3e9615db"
Accept-Ranges: bytes
Content-Length: 6616
Connection: close
Content-Type: text/html
```

mime-counts.000000

A binary file containing the number of files downloaded by mime type.

pageData.000000

A set of files—the output of the mercator.analyzer.html.PageData run-time analyzer—containing an XML-encoded listing of the components of the page.

The elements:

- <URL>--the page's locator
- <PAGEHEADER>--the content of the page <head></head> element
- <JAPPLETCOUNT>--the number of Applets included in the page
- <FORMCOUNT>--the number of forms on the page
- <JSCRIPTCOUNT>--the number of javascripts on the page
- <ILINKS>--the number of links from this page to other pages within the site.
- <ILINK>--the actual internal links
- <ELINKS>--the number of links from this page to external pages.

Example:

```
<?xml version="1.0" encoding="UTF-8"?>
<PAGEPROFILE>
<URL>http://buhariokadigbo.com:80/Press%20Releases/Buhari-
Okadigbo%20Supporters%20in%20United%20States%20Urge%20Nigerian%20Vo\
ters.htm</URL>
<PAGEHEADER>
<head>
<meta content="HTML Tidy, see www.w3.org" name="generator"/>
<meta content="text/html; charset=windows-1252" http-equiv="Content-Type"/>
<meta content="Microsoft FrontPage 4.0" name="GENERATOR"/>
<meta content="FrontPage.Editor.Document" name="ProgId"/>
<title>Buhari/Okadigbo Supporters in United States Urge Nigerian Voters</title>
<meta content="copy-of-straight-edge 000, default" name="Microsoft Theme"/>
```

```

<meta content="tlb, default" name="Microsoft Border"/></head>
</PAGEHEADER>
<JAPPLETCOUNT>0</JAPPLETCOUNT>
<FORMCOUNT>0</FORMCOUNT>
<JSCRIPTCOUNT>0</JSCRIPTCOUNT>
<ILINKS count ="5">
  <ILINK>../images/map_nigeria.gif</ILINK>
  <ILINK>../Links.htm</ILINK>
  <ILINK>mailto:info@buhariokadigbo.com</ILINK>
  <ILINK>mailto:cakukwe@att.net</ILINK>
  <ILINK>mailto:Webmaster@buhariOkadigbo.com</ILINK>
</ILINKS>
<ELINKS count ="1">
  <ELINK> http://www.buhariokadigbo.com/</ELINK>
</ELINKS>
</PAGEPROFILE>

```

robotsLog.000000

The URLs and HTTP status codes for attempted downloads of presumed robots.txt files. In the following example, Mercator tried the URL, but got a "Page not found" return.

Example:

```

http://buhariokadigbo.com:80/robots.txt
Status code 404

```

timings.000000

Mostly more metadata about the crawl. The first line of the example contains metadata about the crawl. The second contains metadata about the page, including HTTP status code, length in bytes, a checksum for the page, and the URL.

Example:

```

start time ID blocked RW lock  fetch process  dns  total
-----
    16  0    2    1   710    550    47   1310

code length  fingerprint  URL
-----
200 21797 1610bc292868e162 http://buhariokadigbo.com:80/

```

Wget

So-called verbose logging from wget records capture time; captured URL and the local URL as it is mirrored on the file system; connection status; content-length and content-type; and download connection speed. It is essentially a record of the request and response history with a minimum of http headers incorporated.

<snip>

```

--10:54:20-- http://dlib.nyu.edu/webarchive
      => `dlib.nyu.edu/webarchive'
Connecting to dlib.nyu.edu:80... connected!
HTTP request sent, awaiting response... 301 Moved Permanently
Location: http://dlib.nyu.edu/webarchive/ [following]
--10:54:20-- http://dlib.nyu.edu/webarchive/

```

```
=> `dlib.nyu.edu/webarchive/index.html'
Connecting to dlib.nyu.edu:80... connected!
HTTP request sent, awaiting response... 200 OK
Length: 3,592 [text/html]
```

```
OK -> ... [100%]
```

```
10:54:20 (3.43 MB/s) - `dlib.nyu.edu/webarchive/index.html' saved [3592/3592]
```

```
Loading robots.txt; please ignore errors.
--10:54:21-- http://dlib.nyu.edu/robots.txt
=> `dlib.nyu.edu/robots.txt'
Connecting to dlib.nyu.edu:80... connected!
HTTP request sent, awaiting response... 200 OK
Length: 168 [text/plain]
```

```
OK -> [100%]
```

</snip>

```
10:54:21 (164.06 KB/s) - `dlib.nyu.edu/robots.txt' saved [168/168]
```

```
--10:54:21-- http://dlib.nyu.edu/webarchive/prototypes.html
=> `dlib.nyu.edu/webarchive/prototypes.html'
Connecting to dlib.nyu.edu:80... connected!
HTTP request sent, awaiting response... 200 OK
Length: 4,308 [text/html]
```

```
OK -> .... [100%]
```

```
10:54:21 (4.11 MB/s) - `dlib.nyu.edu/webarchive/prototypes.html' saved [4308/4308]
```

```
--10:54:21-- http://dlib.nyu.edu/webarchive/index.html
=> `dlib.nyu.edu/webarchive/index.html'
Connecting to dlib.nyu.edu:80... connected!
HTTP request sent, awaiting response... 200 OK
Length: 3,592 [text/html]
```

```
OK -> ... [100%]
```

</snip>

Linklint

Linklint is essentially a link checker and not a mirroring crawler per se, but it can crawl a site and produce a panoply of useful reports that monitor the state of the link structure, as well as giving a summary of the files contained in the site.

The logs for a single HTTrack capture of one of the Nigerian Election sites can be found here:
<http://dlib.nyu.edu/webarchive/linklintlogs/socialistnigeria/index.html>

A. summary.txt

This log reports on the overall structure and health of the site - the total number of directories; whether there is a default index; the total number of files and its breakdown between HTML, image and other; external links e.g. mailtos that should be disabled; missing links, missing internal anchors.

It records the root of the site; date of capture and software version at the top of the file, and specifies errors at the bottom.

<snip>

```
file: summary.txt
root: /www.socialistnigeria.org
date: Sun, 27 Apr 2003 11:43:17 (local)
Linklint version: 2.3.5
```

```
Linklint found 102 files in 16 directories and checked 91 html files.
There were no missing files. No files had broken links.
1 error, no warnings.
```

```
found 16 directories with files
found 1 default index
found 90 html files
found 2 image files
found 9 other files
found 1 http link
found 3 mailto links
found 6 named anchors
----- 1 action skipped
ERROR    1 missing named anchor
```

</snip>

B. **log.txt**

Records the progress of the crawl.

C. **dir.txt**

Gives a list of directories

D. **file.txt**

Lists files by type

<http://dlib.nyu.edu/webarchive/linklintlogs/socialistnigeria/file.htm>

E. **fileX.txt**

Records cross-referencing/cross-linking amongst the files

<http://dlib.nyu.edu/webarchive/linklintlogs/socialistnigeria/fileX.htm>

F. **fileF.txt**

Records the link structure of the Web site by recording the forward links found in every file.

<http://dlib.nyu.edu/webarchive/linklintlogs/socialistnigeria/fileF.htm>

G. **remote.txt** and remoteX.txt

Record the external http links found and the files that contain them.

H. **anchors.txt** and anchorsX.txt

Record named anchors and where they are found.

I. **action.txt** and actionX.txt

Record actions that were skipped and where those action links occurred. This involves form input for the most part.

In the following case the link is to an external site and to a perl script:

<http://dlib.nyu.edu/webarchive/linklintlogs/socialistnigeria/action.htm>

J. **errorA.txt** and errorAX.txt

Records errors in named anchors and the files in which they occur.

N.B. The following occur in other mirrors:

K. **warn.txt**

Various warnings show up here. In this case no single index.html file was found (since the homepage is a frameset) :

<http://dlib.nyu.edu/webarchive/linklintlogs/allidemoUK/warn.htm>

L. **imgmap.txt** and imgmapX.txt

Records named image maps and the files in which they occur.

<http://dlib.nyu.edu/webarchive/linklintlogs/peoplesmandate/imgmap.htm>

Recommendations

A combination of focused IA crawling, Mercator or an application such as PANDAS wrapped around HTTrack paired with Linklint would appear to provide a sufficient amount of capture metadata for preservation. A further application e.g. ImageMagick for images, should be considered to extract more specific preservation metadata e.g. pixels wide and pixels high; bits per sample; compression and so on.

The mechanism for extracting the metadata from these logs could be as simple as a series of perl scripts. Any scripting language that can do system calls, recurse through directories and files, read in files in binmode, and process text will suffice for this task. These scripts would output SQL statements or CSV load files that can be dumped into a database for further processing. In progress is a series of perl scripts that can make a series of SQL queries against a Web site's metadata in the database and create the dmdSec, amdSec, fileSec, structMap and structLink for METS description/encapsulation of a site.

APPENDIX 21

Feasibility of Populating a METS File from an IA SIP (.arc + .dat)

This appendix seeks to demonstrate how METS (the Metadata Encoding and Transmission Standard)ⁱ is uniquely suited to address several challenges specific to archiving web sites, e.g. articulating the structure and boundaries of a site as well as the complexities of interrelations within it; and the management of descriptive, preservation, and structural metadata necessary both to provide continued access to the content and to ensure that the repository can continue to rematerialize a digital object whose medium is particularly volatile and ephemeral, and whose components will require frequent refreshing or migration for long-term preservation.ⁱⁱ

Due primarily to its flexibility, extensibility and interoperability, XML has become the recognized *lingua franca* of data (and metadata) exchange in the digital realm. It is also gradually becoming the currency of digital object management in OAIS-compliant repositories. The Nordic Web Archive's NWA Toolsetⁱⁱⁱ opened the way to exposing archived web site metadata in XML in the form of their own NWA Document Format schema,^{iv} which comprises descriptive and preservation metadata for a single web page along with a list of links parsed from the page. This is a step in the right direction, but it arguably does not go far enough on two counts: a) it is a proprietary schema, and b) it catalogs what is a discrete sub-object, and does not encompass the structure of the web site as a complex, articulated digital object along with the interrelations between its components.

Some of the weaknesses in the harnessing of the NEDLIB harvester to the NWA Access Module (using FAST against indexed NWA Document Format documents) were revealed in the final report of the Danish netarkivet.dk web-archiving project.^v The netarkivet.dk project is of particular interest and import to the CRL project because it was a pilot project for archiving political web sites during the November 2001 Danish County and District Elections. Indexing was done at the page-level, and the lack of structural metadata delineating the web site as a whole led to the inevitable result that queries returned a collection of URLs (sometimes thousands) without any clear indication where to find the home page for any given page and thus contextualize it. METS is fully poised to overcome these stumbling blocks.

As a syntax METS is best equipped to address the complexities of describing and allowing users to navigate the structure of a web site through the interworkings of the structMap, structLink and fileGrp sections. Within the all-important structMap, elements such as <par> for parallel elements properly address the synchronic nature of some linking, (e.g. the rhizomic structure of embedded SRCs such as inline graphics that occur across a whole set of pages in one web site).

METS is also equipped to handle the complexities of managing the recording of any changes undergone by an object, both as modified by the creator and as altered by the archival body. Different versions of an object modified by the creator can be managed within the <fileGrp>, while the <digiprov> section, a subset of <amdSec> in METS, tracks the digital provenance of an object as it undergoes any alteration in the course of its preservation. This might include refreshing of bits and/or migration to another, standardized, format.

As part of the NDIP Plan the Library of Congress is vigorously examining the possibility of using METS to describe web sites;^{vi} Morgan Cundiff has laid the groundwork for such an examination in his 2002 ECDL presentation.^{vii} In the meantime, the CRL Political Communications Web Archiving project,^{viii} in partnership with the Internet Archive, is poised to evaluate an implementation of METS as a means to both management of and access to web sites in their entirety that have been archived in the Alexa/Internet Archive .arc format with an accompanying .dat metadata file.^{ix}

The following analysis will make use of the homepage of the site owned by Le Front Social (<http://perso.magic.fr/nac/>), as archived on 13 June 2000, to explore the feasibility of extracting descriptive, administrative and structural metadata out of Internet Archive SIP (consisting of .arc and .dat files for that page) into a METS wrapper that will serve as an AIP and/or a DIP.

When the homepage (<http://perso.magic.fr/nac/index.htm>) for the site was archived, it contained an embedded Flash application and an href link to enter the site at index2.html. Pages that embed Flash are recognized as problematic for web archiving, especially when they contain embedded links; in addition, their sheer size can be

prohibitive in cases where storage space is an issue. This archived version is not available from the Wayback Machine (probably because of a file size limit imposed on the crawl); in all other available Wayback Machine archives of the site the opening Flash page has been superseded by a modified version of what has been archived in this SIP package as index2.html.

1.0 The Internet Archive Files

An .arc file is essentially an aggregation of captured HTML and associated server-delivered and crawler-generated metadata for each page bundled into a 100 MB archive file.^x A generic Alexa .arc file is entirely promiscuous and random, according to where that particular crawler wandered in that snapshot, and how much could be packed into a 100 MB file. However, the .arc files that the CRL Project is receiving will have undergone one further level of handling, to collect each of four region's URLs into its own .arc file(s). An .arc file is complemented by a .dat (metadata) and a .cdx (index) file. There is also a complete .cdx index for all the .arcs residing on a single Internet Archive hard disk.

How does the .arc/.dat package stack up as an information package that could be used as an AIP in an OAIS-compliant archive? The major weakness of the .arc format that may consign it to serving as a SIP that will be transformed into an AIP and DIP, to my mind, is that it is not expressed as XML. It also offers no readily discernible structural metadata for a web site, although the links originating in a single page are enumerated in the .dat file; in order to fully process link structure, however, the HTML for archived pages in the .arc file itself will have to be processed.

1.1 index.htm

1.1.1 The .dat file for index.htm

Of immediate interest is the IA .dat file's helpful breakdown of parsed links into href links and embedded srcs, respectively, for the two linked items; this distinction will facilitate processing of the <structMap> and <structLink> sub-elements. Two types of md5 checksum are recorded in a MIX <checksum> element for use in an integrity check. Also note that the metaparser seems to have stripped out the page title's eacute character, which was not rendered by an HTML entity in the original page.

To fully decode the field names for the following file, see the key to the .dat fields in [Appendix 20](#) above.

```
http://perso.magic.fr:80/nac/index.htm 195.115.16.3 20000613200345 alexa/dat 195
m text/html
s 200
c d9ca1a19f87795931a8fcb77927056cd
k 0d22933b758ec5497fdb39ee7d06eb20
v 821541
V 272047
n 832
t Vive la r evolution!
l perso.magic.fr/nac/index2.html
y perso.magic.fr/nac/sw.swf
```

1.1.2 The .arc file for index.htm

The .arc file consists of http headers sent by the server, coupled with the HTML of the page itself, or the binary of a multimedia or executable file.

```
http://perso.magic.fr:80/nac/index.htm 195.115.16.3 20000613200345 text/html 832
HTTP/1.1 200 OK
Date: Tue, 13 Jun 2000 20:03:45 GMT
```

```

Server: MOL-UNIX/AP-FP-PHP-PERL-FCGI-XML/29022000/DED
Last-Modified: Fri, 31 Mar 2000 00:13:07 GMT
ETag: "6e291-22f-38e3ed93"
Accept-Ranges: bytes
Content-Length: 559
Connection: close
Content-Type: text/html

```

```

<HTML>
<HEAD>

  <TITLE>Vive la révolution!</TITLE>
  <META HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=iso-8859-1">
</HEAD>
<BODY BGCOLOR="#ffffff" LINK="#990000" VLINK="#990000">

  <P><CENTER><B><FONT FACE="Courier New, Courier, mono"><A HREF="index2.html"
  TARGET="">Cliquez ici pour le site!</A></FONT></B></CENTER></P>

  <P><CENTER><FONT COLOR="#330000"><font></FONT></font><EMBED SRC="sw.swf"
  pluginspage="http://www.macromedia.com/shockwave/download/" WIDTH="700"
  HEIGHT="510" loop="false" ALIGN="BOTTOM"></EMBED>
  </CENTER>

</BODY>
</HTML>

```

2.0 Possible METS output

In the following section I have manually generated discrete METS descriptive, administrative and structural elements for this simple page and its two linked items, another webpage and a Flash binary that was not archived.

In order to automate the extraction of metadata from these files perl or some other scripting/programming language could be used to parse the .arc and .dat and hold the metadata for each object in a data structure, from which it could populate a METS file. Alternatively, metadata could be parsed out of the .arc and .dat files into a database, from which METS would be extracted.

2.1 index.htm

2.1.1 Descriptive Metadata <dmdSec> for the page:

At this point in time I envision wrapping either DC or MODS-Lite for descriptive metadata at the page level, with a full MODS description at the web site level. Ideally extraction of the page-level <dmdSec> information can be entirely automated, while the top-level description for the entire web site will require some human analysis and cataloguing in addition to what can be automatically extracted.

For the entire web site:

```

<METS:dmdSec ID="DM01">
  <METS:mdWrap MDTYPE="MODS">
    <METS:xmlData>
  <mods>
  <titleInfo>
    <title>Front Social Web Site</title>
  </titleInfo>
  <titleInfo type="alternative">
    <title>Front Social</title>
  </titleInfo>
  <name type="corporate">
    <namePart>Front Social</namePart>
  </name>

```

```

    <genre>Web site</genre>
    <originInfo>
    <dateCaptured point="start" encoding="iso8601">20010617</dateCaptured>
    <dateCaptured point="end" encoding="iso8601">20020402</dateCaptured>
    </originInfo>
    <language authority="iso639-2b">fre</language>
    <physicalDescription>
    <internetMediaType>text/html</internetMediaType>
    <internetMediaType>image/jpeg</internetMediaType>
    </physicalDescription>
    <abstract>Web site for the online publication Front Social/Triple Oppression, archiving publications from September
    1995 to September 2001.</abstract>
    <subject>
    <topic>Marxism</topic>
    <geographic>France</geographic>
    </subject>
    <subject>
    <topic>Maoism</topic>
    <geographic>France</geographic>
    </subject>
    <subject>
    <topic>Socialism</topic>
    <geographic>France</geographic>
    </subject>
    <subject>
    <topic>Communist Party</topic>
    <geographic>France</geographic>
    </subject>
    <subject>
    <topic>Autonomy Movement</topic>
    <geographic>France</geographic>
    </subject>
    <relatedItem type="host">
    <titleInfo>
    <title>CRL Political Web Archiving Project</title>
    </titleInfo>
    <identifier type="uri">http://www.crl.edu/content/PolitWeb.htm</identifier>
    </relatedItem>
    <identifier type="uri" displayLabel="Active site (if available)">http://perso.magic.fr/nac</identifier>
    <identifier type="uri" displayLabel="Archived site">http://crawl04.archive.org/crl/*/http://perso.magic.fr/nac</identifier>
    </mods>
  </METS:xmlData>
</METS:mdWrap>
</METS:dmdSec>

```

For the index.htm page itself:

```

<METS:dmdSec ID="DM1">
  <METS:mdWrap MDTYPE="DC">
    <METS:xmlData>
      <dc:title>Vive la révolution!</dc:title>
      <dc:identifier> http://perso.magic.fr:80/nac/index.htm </dc:identifier>
      <dc:extent>559</dc:extent>
      <dc:date>2003-03-31T00:13:07</dc:date>
    </METS:xmlData>
  </METS:mdWrap>
</METS:dmdSec>

```

The title was more completely derived from the HTML page itself as archived in the .arc file than from the .dat, which lacked the acute character. The URL/identifier could be parsed from either the .dat or the .arc header for the file. The extent element can be taken from the .dat or the http header in the .arc. The date, assuming that the last modified date is the desired value here, will need to be programmatically normalized, but can be extracted

from the Last-Modified-Date field in the http headers. Occasionally a page will include a <meta> tag containing the original creation date of the file.

2.1.2 Administrative <admSec>/<techMD>

At the page level very little technical information is available from the combination of .arc and .dat. The <charset> would have to be extracted from the archived HTML's Content-Type <meta> tag if extant, which, though required for validation, is not apt to be present in the majority of pages captured. The language and markup language fields could be analyzed programmatically, especially the latter value.

```
<METS:techMD ID="ADM1">
  <METS:mdWrap MDTYPE="OTHER">
    <METS:xmlData>
      <textmd:textMD>
        <textmd:charset>iso-8859-1</textmd:charset>
        <textmd:language>fr</textmd:language>
        <textmd:markup_language>HTML</textmd:markup_language>
        <textmd:textNote>index.htm</textmd:textNote>
      </textmd:textMD>
    </METS:xmlData>
  </METS:mdWrap>
</METS:techMD>
```

2.1.3 File Information <fileSec>

A <fileSec> can be programmatically generated from the links parsed into the .dat file. MIMETYPE and CREATED attributes should be derived from the http headers in the .arc file, along with the xlink:href value. Programming challenges might arise in expressing different <fileGrp> configurations.

```
<METS:fileSec>
  <METS:fileGrp>
    <METS:file ID="FID1" MIMETYPE="text/html" SEQ="1" CREATED="2003-03-31T00:13:07" ADMID="ADM1">
      <METS:FLocat LOCTYPE="URL" xlink:href="http://perso.magic.fr:80/nac/index.htm"></METS:FLocat>
    </METS:file>

    <METS:file ID="FID2" MIMETYPE="" application/x-shockwave-flash" SEQ="2" CREATED="2003-03-31T00:13:07"
    ADMID="ADM2">
      <METS:FLocat LOCTYPE="URL" xlink:href="http://perso.magic.fr:80/nac/sw.swf"></METS:FLocat>
    </METS:file>

    <METS:file ID="FID3" MIMETYPE="text/html" SEQ="3" CREATED="2003-06-06T23:54:30" ADMID="ADM3">
      <METS:FLocat LOCTYPE="URL" xlink:href="http://perso.magic.fr:80/nac/index2.html"></METS:FLocat>
    </METS:file>
  </METS:fileGrp>
</METS:fileSec>
```

2.1.4 Structural Information <structMap>

The <div> structure that can be derived from the homepage alone is of course only a snippet of the full <structMap> for a site, but will serve for the sake of a preliminary demonstration. In the structure that would evolve out of a METS extraction from a complete web site's metadata, the index.html page (or alternatively the frameset of a frame-based homepage) would stand at the root.

In the following case the embedded .swf file, like other embedded or inline files meant to be presented simultaneously with the page (e.g. images) is treated as a parallel element <par> in the index.htm file, while the href link to index2.html is a nested <div> under the root page <div>. We can think of the distinction in terms of synchronic vs diachronic linking.

Recursively parsing out the <div> and embedded link structures from a database or data structure extracted from the aggregate of files bundled into the .arc will be one of the more interesting programming challenges of this project.

The actual byte offset location in the HTML page of the links will have to be generated using perl or java to fill out the BEGIN and END attribute values (these values are not real).

```
<METS:structMap>
  <METS:div ID="PAGE1" ORDER="1" LABEL="http://perso.magic.fr:80/nac/index.htm" DMDID="DM1">
    <METS:fptr>
      <METS:par>
        <METS:area FILEID="FID3" BEGIN="100" END="120" BETYPE="BYTE"/>
      </METS:par>
    </METS:fptr>
    <METS:div ID="LINK1" ORDER="1" LABEL="Link to Vive la Guerre populaire au P&eacute;rou, au N&eacute;pal, et
partout ailleurs! Vive le marxisme-l&eacute;onisme-mao&iuml;sme!">
  <METS:fptr><METS:area FILEID="FID1" BEGIN="100" END="120" BETYPE="BYTE"/></METS:fptr>
</METS:div>
[nested <divs> follow for each of the remaining pages under index.htm as the root for the site]
</METS:div>
</METS:structMap>
```

2.1.5 Linking Information <structLink>

Again, the links available from the homepage make up a small subset of the full <structLink> for this site. In this case the link to the second index page is recorded, but the embedded SRC to the Flash page is not; the latter is a parallel file and not a hyperlink. The same would hold true for embedded graphics or image files

The HTML embedded in the .arc file will have to be processed by perl or some other language in order to extract the link structure.

```
<METS:structLink>
  <METS:smLink from="LINK1" to="PAGE3" xlink:title="Cliquez ici pour le site!"/>
  [href links to other pages follow]
</METS:structLink>
```

3. Conclusion

In order to transform the Internet Archive .arc and .dat files into METS objects for use as an AIP or DIP a fair amount of complex data processing will have to be undertaken. I can envision a scripting language such as Perl or Ruby as an adequate tool for this task, with a little help from Unix compression/decompression tools, whether it involves reading the metadata for an object from the .dat and .arc files into a data structure or into a database and then extracting METS from there.

Unless a (highly unlikely) metadata exchange is negotiated with creators/owners of web sites, perhaps along with permissions, projects involved in web archiving have access to as much descriptive and technical metadata for the harvested pages and multimedia objects as they can wrangle out of their archived files. The Internet Archive SIP is adequate for automated extraction of a fairly good capture metadata set and a rather poor descriptive metadata set (barely enough for simple discovery).

This descriptive and even technical information may occasionally be misleading or wrong, as in the case of much-abused description and keyword <meta> tags, whose import in the end may only be archival, or in the case of wholesale pirated <meta> tags, as demonstrated by index2.html from the site I've chosen to serve as a model for this study. In the end cataloguers and subject specialists will surely be called upon to provide an objective, analytical description of a web site for the top-level record, be it METS pointing to or wrapping MODS or DC. Whether a repository or service such as Webarchivist.org has the capacity to produce a human-vetted descriptive record for each page is a question of personnel and funding in an already incredibly expensive enterprise.

Whereas archived HTML pages along with their http headers are an easily parsed source of metadata, the processing of multimedia files and other binaries will be a particular challenge. In the case of images, where IPTC headers occur there is a Perl Module available for parsing out information; otherwise ImageMagick appears to be the best solution for extracting information such as geometry, resolution, compression, bits per sample, and so on. Images are only a fraction of the multimedia and binary files that will be archived for any given website; the rest are the subject of another study.

APPENDIX 22

METS Template for a website

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- aniagolu_20030417.xml
Prepared for the CRL Political Web Archiving Project by Leslie Myrick,
Digital Library Team, New York University Libraries
leslie.myrick@nyu.edu
2003 JULY 30
Version 1.0
updated 2004 April 29 to reflect new MODS version
-->

<!-- This file was hand-generated using emacs, perl and regex. Sources included an HTTrack
mirror
and a Linklint fileF.txt file (for link structure). Some of the values are not true values,
e.g. the byte offsets in the HTML page. MODS generation is indebted to the model ulmer.xml
file posted on the MINERVA site. -->

<METS:mets
  xmlns:METS="http://www.loc.gov/METS/"
  xmlns:mix="http://www.loc.gov/mix/"
  xmlns:mods="http://www.loc.gov/mods/v3"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:xlink="http://www.w3.org/TR/xlink"
  xmlns:textmd="http://dlib.nyu.edu/METS/textMD"
  xsi:schemaLocation="http://www.loc.gov/METS/ http://www.loc.gov/standards/mets/mets.xsd
http://www.loc.gov/mix/ http://www.loc.gov/standards/mix/mix.xsd
http://www.loc.gov/mods/v3 http://www.loc.gov/standards/mods/v3/mods-3-0.xsd
http://purl.org/dc/elements/1.1/ http://dublincore.org/schemas/xmls/simpledc20021212.xsd
http://dlib.nyu.edu/METS1/textMD file:/C:/eadcb/METSFiles/textmd.xsd
http://dlib.nyu.edu/METS/textMD file:/C:/eadcb/METSFiles/textmd.xsd
http://www.w3.org/TR/xlink/ http://www.loc.gov/standards/mets/xlink.xsd"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  OBJID="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/aniagolu_20030417"
  TYPE="Website">
  <METS:metsHdr CREATEDATE="2003-04-17T12:46:40" LASTMODDATE="2003-07-30T13:46:40">
    <METS:agent ROLE="CREATOR">
      <METS:name>LESLIE</METS:name>
    </METS:agent>
  </METS:metsHdr>
  <METS:dmdSec ID="DM01">
    <METS:mdWrap MDTYPE="MODS">
      <METS:xmlData>
        <mods:mods>
          <mods:titleInfo>
            <mods:title>Loretta Aniagolu Web Site</mods:title>
          </mods:titleInfo>
          <mods:name type="personal">
            <mods:namePart>Aniagolu, Loretta</mods:namePart>
          </mods:name>
          <mods:genre>Web site</mods:genre>
          <mods:originInfo>
            <mods:dateCaptured encoding="iso8601">20030417</mods:dateCaptured>
          </mods:originInfo>
          <mods:language><mods:languageTerm authority="iso639-
2b">eng</mods:languageTerm></mods:language>
          <mods:physicalDescription>
            <mods:internetMediaType>text/html</mods:internetMediaType>
            <mods:internetMediaType>image/jpeg</mods:internetMediaType>
            <mods:internetMediaType>image/gif</mods:internetMediaType>
          </mods:physicalDescription>
        </mods:mods>
      </METS:xmlData>
    </METS:mdWrap>
  </METS:dmdSec>
</METS:mets>
```

```

candidate
    <mods:abstract>Web site promoting the candidacy of Loretta Aniagolu, NCP
    for governor of the Enugu State in the April 2003 Nigerian Election.
    Includes biography page, photo gallery, agenda, selected
links</mods:abstract>
    <mods:subject>
        <mods:topic>Elections</mods:topic>
        <mods:geographic>Africa</mods:geographic>
        <mods:geographic>Enugu State</mods:geographic>
    </mods:subject>
    <mods:relatedItem type="host">
        <mods:titleInfo>
            <mods:title>CRL Political Web Archiving Project</mods:title>
        </mods:titleInfo>
        <mods:identifier
type="uri">http://www.crl.edu/content/PolitWeb.htm</mods:identifier>
        </mods:relatedItem>
        <mods:identifier displayLabel="Archived site"
type="uri">http://ia00644.archive.org/nigeria/*/http://www.aniagolu.org/</mods:identifier>
        </mods:mods>
    </METS:xmlData>
</METS:mdWrap>
</METS:dmdSec>
<METS:dmdSec ID="DM1">
    <METS:mdWrap MDTYPE="MODS">
        <METS:xmlData>
            <mods:mods>
                <mods:titleInfo><mods:title>Welcome to my
Website</mods:title></mods:titleInfo>
                <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/index.html</mods:identifi
er>
                <mods:physicalDescription><mods:extent
type="filesize">35710</mods:extent></mods:physicalDescription>
                <mods:originInfo><mods:dateCaptured>2003-04-
17T18:58:16</mods:dateCaptured></mods:originInfo>
                </mods:mods>
            </METS:xmlData>
        </METS:mdWrap>
    </METS:dmdSec>
<METS:dmdSec ID="DM2">
    <METS:mdWrap MDTYPE="MODS">
        <METS:xmlData>
            <mods:mods>
                <mods:titleInfo><mods:title>Who Is She?</mods:title></mods:titleInfo>
                <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/aboutme.htm</mods:identif
ier>
                <mods:physicalDescription><mods:extent
type="filesize">16927</mods:extent></mods:physicalDescription>
                <mods:originInfo><mods:dateCaptured>2003-04-
17T14:32:26</mods:dateCaptured></mods:originInfo>
                </mods:mods>
            </METS:xmlData>
        </METS:mdWrap>
    </METS:dmdSec>
<METS:dmdSec ID="DM3">
    <METS:mdWrap MDTYPE="MODS">
        <METS:xmlData>
            <mods:mods>
                <mods:titleInfo><mods:title>Agenda, not Gender
...</mods:title></mods:titleInfo>

```



```

                <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/agenda.htm</mods:identifi
er>
                <mods:physicalDescription><mods:extent
type="filesize">26726</mods:extent></mods:physicalDescription>
                <mods:originInfo><mods:dateCaptured>2003-04-
17T14:32:26</mods:dateCaptured></mods:originInfo>
                </mods:mods>
            </METS:xmlData>
        </METS:mdWrap>
    </METS:dmdSec>
    <METS:dmdSec ID="DM4">
        <METS:mdWrap MDTYPE="MODS">
            <METS:xmlData>
                <mods:mods>
                    <mods:titleInfo><mods:title>Favorites</mods:title></mods:titleInfo>
                    <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/favorite.htm</mods:identi
fier>
                    <mods:physicalDescription><mods:extent
type="filesize">19937</mods:extent></mods:physicalDescription>
                    <mods:originInfo><mods:dateCaptured>2003-04-
17T14:32:26</mods:dateCaptured></mods:originInfo>
                    </mods:mods>
                </METS:xmlData>
            </METS:mdWrap>
        </METS:dmdSec>
        <METS:dmdSec ID="DM5">
            <METS:mdWrap MDTYPE="MODS">
                <METS:xmlData>
                    <mods:mods>
                        <mods:titleInfo><mods:title>Photo
Gallery</mods:title></mods:titleInfo>
                        <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/photo.htm</mods:identifie
r>
                        <mods:physicalDescription><mods:extent
type="filesize">15120</mods:extent></mods:physicalDescription>
                        <mods:originInfo><mods:dateCaptured>2003-04-
17T14:32:26</mods:dateCaptured></mods:originInfo>
                        </mods:mods>
                    </METS:xmlData>
                </METS:mdWrap>
            </METS:dmdSec>
            <METS:dmdSec ID="DM6">
                <METS:mdWrap MDTYPE="MODS">
                    <METS:xmlData>
                        <mods:mods>
                            <mods:titleInfo><mods:title>Feedback</mods:title></mods:titleInfo>
                            <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/feedback.htm</mods:identi
fier>
                            <mods:physicalDescription><mods:extent
type="filesize">11783</mods:extent></mods:physicalDescription>
                            <mods:originInfo><mods:dateCaptured>2003-04-
17T14:32:26</mods:dateCaptured></mods:originInfo>
                            </mods:mods>
                        </METS:xmlData>
                    </METS:mdWrap>
                </METS:dmdSec>
                <METS:dmdSec ID="DM7">
                    <METS:mdWrap MDTYPE="MODS">
                        <METS:xmlData>

```

```

        <mods:mods>
            <mods:titleInfo><mods:title>[counter]</mods:title></mods:titleInfo>
            <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_vti_bin/fpcount.exe/inde
x4062.html</mods:identifier>
            <mods:physicalDescription><mods:extent
type="filesize">208</mods:extent></mods:physicalDescription>
            <mods:originInfo><mods:dateCaptured>2003-04-
19T10:36:26</mods:dateCaptured></mods:originInfo>
            </mods:mods>
        </METS:xmlData>
    </METS:mdWrap>
</METS:dmdSec>
<METS:dmdSec ID="DM8">
    <METS:mdWrap MDTYPE="MODS">
        <METS:xmlData>
            <mods:mods>

<mods:titleInfo><mods:title>home_cmp_network010_vbtn_p.gif</mods:title></mods:titleInfo>
            <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/home_cmp_network
010_vbtn_p.gif</mods:identifier>
            <mods:physicalDescription><mods:extent
type="filesize">212</mods:extent></mods:physicalDescription>
            <mods:originInfo><mods:dateCaptured>2003-04-
17T17:18:00</mods:dateCaptured></mods:originInfo>
            </mods:mods>
        </METS:xmlData>
    </METS:mdWrap>
</METS:dmdSec>
<METS:dmdSec ID="DM9">
    <METS:mdWrap MDTYPE="MODS">
        <METS:xmlData>
            <mods:mods>

<mods:titleInfo><mods:title>home_cmp_network010_vbtn_a.gif</mods:title></mods:titleInfo>
            <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/home_cmp_network
010_vbtn_a.gif</mods:identifier>
            <mods:physicalDescription><mods:extent
type="filesize">218</mods:extent></mods:physicalDescription>
            <mods:originInfo><mods:dateCaptured>2003-04-
17T17:18:00</mods:dateCaptured></mods:originInfo>
            </mods:mods>
        </METS:xmlData>
    </METS:mdWrap>
</METS:dmdSec>
<METS:dmdSec ID="DM10">
    <METS:mdWrap MDTYPE="MODS">
        <METS:xmlData>
            <mods:mods>

<mods:titleInfo><mods:title>aboutme.htm_cmp_network010_vbtn.gif</mods:title></mods:titleInfo>
            <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/aboutme.htm_cmp_
network010_vbtn.gif</mods:identifier>
            <mods:physicalDescription><mods:extent
type="filesize">204</mods:extent></mods:physicalDescription>
            <mods:originInfo><mods:dateCaptured>2003-04-
17T17:18:00</mods:dateCaptured></mods:originInfo>
            </mods:mods>
        </METS:xmlData>
    </METS:mdWrap>

```

```

</METS:dmdSec>
<METS:dmdSec ID="DM11">
  <METS:mdWrap MDTYPE="MODS">
    <METS:xmlData>
      <mods:mods>

<mods:titleInfo><mods:title>aboutme.htm_cmp_network010_vbtn_a.gif</mods:title></mods:titleInfo>
<mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/aboutme.htm_cmp_network010_vbtn_a.gif</mods:identifier>
  <mods:physicalDescription><mods:extent
type="filesize">251</mods:extent></mods:physicalDescription>
  <mods:originInfo><mods:dateCaptured>2003-04-17T17:18:00</mods:dateCaptured></mods:originInfo>
  </mods:mods>
    </METS:xmlData>
  </METS:mdWrap>
</METS:dmdSec>
<METS:dmdSec ID="DM12">
  <METS:mdWrap MDTYPE="MODS">
    <METS:xmlData>
      <mods:mods>

<mods:titleInfo><mods:title>agenda.htm_cmp_network010_vbtn.gif</mods:title></mods:titleInfo>
  <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/agenda.htm_cmp_network010_vbtn.gif</mods:identifier>
  <mods:physicalDescription><mods:extent
type="filesize">248</mods:extent></mods:physicalDescription>
  <mods:originInfo><mods:dateCaptured>2003-04-17T17:18:00</mods:dateCaptured></mods:originInfo>
  </mods:mods>
    </METS:xmlData>
  </METS:mdWrap>
</METS:dmdSec>
<METS:dmdSec ID="DM13">
  <METS:mdWrap MDTYPE="MODS">
    <METS:xmlData>
      <mods:mods>

<mods:titleInfo><mods:title>agenda.htm_cmp_network010_vbtn_a.gif</mods:title></mods:titleInfo>
  <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/agenda.htm_cmp_network010_vbtn_a.gif</mods:identifier>
  <mods:physicalDescription><mods:extent
type="filesize">303</mods:extent></mods:physicalDescription>
  <mods:originInfo><mods:dateCaptured>2003-04-17T17:18:00</mods:dateCaptured></mods:originInfo>
  </mods:mods>
    </METS:xmlData>
  </METS:mdWrap>
</METS:dmdSec>
<METS:dmdSec ID="DM14">
  <METS:mdWrap MDTYPE="MODS">
    <METS:xmlData>
      <mods:mods>

<mods:titleInfo><mods:title>favorite.htm_cmp_network010_vbtn.gif</mods:title></mods:titleInfo>
  </mods:mods>
    </METS:xmlData>
  </METS:mdWrap>
</METS:dmdSec>

```

```

                <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/favorite.htm_cmp_network010_vbtn.gif</mods:identifier>
                <mods:physicalDescription><mods:extent
type="filesize">191</mods:extent></mods:physicalDescription>
                <mods:originInfo><mods:dateCaptured>2003-04-
17T17:18:00</mods:dateCaptured></mods:originInfo>
                </mods:mods>
            </METS:xmlData>
        </METS:mdWrap>
    </METS:dmdSec>
    <METS:dmdSec ID="DM15">
        <METS:mdWrap MDTYPE="MODS">
            <METS:xmlData>
                <mods:mods>

```

```

<mods:titleInfo><mods:title>favorite.htm_cmp_network010_vbtn_a.gif</mods:title></mods:titleInfo>

```

```

                <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/favorite.htm_cmp_network010_vbtn_a.gif</mods:identifier>
                <mods:physicalDescription><mods:extent
type="filesize">239</mods:extent></mods:physicalDescription>
                <mods:originInfo><mods:dateCaptured>2003-04-
17T17:18:00</mods:dateCaptured></mods:originInfo>
                </mods:mods>
            </METS:xmlData>
        </METS:mdWrap>
    </METS:dmdSec>
    <METS:dmdSec ID="DM16">
        <METS:mdWrap MDTYPE="MODS">
            <METS:xmlData>
                <mods:mods>

```

```

<mods:titleInfo><mods:title>photo.htm_cmp_network010_vbtn.gif</mods:title></mods:titleInfo>

```

```

                <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/photo.htm_cmp_network010_vbtn.gif</mods:identifier>
                <mods:physicalDescription><mods:extent
type="filesize">216</mods:extent></mods:physicalDescription>
                <mods:originInfo><mods:dateCaptured>2003-04-
17T17:18:00</mods:dateCaptured></mods:originInfo>
                </mods:mods>
            </METS:xmlData>
        </METS:mdWrap>
    </METS:dmdSec>
    <METS:dmdSec ID="DM17">
        <METS:mdWrap MDTYPE="MODS">
            <METS:xmlData>
                <mods:mods>

```

```

<mods:titleInfo><mods:title>photo.htm_cmp_network010_vbtn_a.gif</mods:title></mods:titleInfo>

```

```

                <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/photo.htm_cmp_network010_vbtn_a.gif</mods:identifier>
                <mods:physicalDescription><mods:extent
type="filesize">268</mods:extent></mods:physicalDescription>
                <mods:originInfo><mods:dateCaptured>2003-04-
17T17:18:00</mods:dateCaptured></mods:originInfo>
                </mods:mods>
            </METS:xmlData>
        </METS:mdWrap>
    </METS:dmdSec>

```

```

<METS:dmdSec ID="DM18">
  <METS:mdWrap MDTYPE="MODS">
    <METS:xmlData>
      <mods:mods>

<mods:titleInfo><mods:title>feedback.htm_cmp_network010_vbtn.gif</mods:title></mods:titleInfo>
>
      <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/feedback.htm_cmp_network010_vbtn.gif</mods:identifier>
      <mods:physicalDescription><mods:extent
type="filesize">194</mods:extent></mods:physicalDescription>
      <mods:originInfo><mods:dateCaptured>2003-04-17T17:18:00</mods:dateCaptured></mods:originInfo>
      </mods:mods>
    </METS:xmlData>
  </METS:mdWrap>
</METS:dmdSec>
<METS:dmdSec ID="DM19">
  <METS:mdWrap MDTYPE="MODS">
    <METS:xmlData>
      <mods:mods>

<mods:titleInfo><mods:title>feedback.htm_cmp_network010_vbtn_a.gif</mods:title></mods:titleInfo>
fo>
      <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/feedback.htm_cmp_network010_vbtn_a.gif</mods:identifier>
      <mods:physicalDescription><mods:extent
type="filesize">240</mods:extent></mods:physicalDescription>
      <mods:originInfo><mods:dateCaptured>2003-04-17T17:18:00</mods:dateCaptured></mods:originInfo>
      </mods:mods>
    </METS:xmlData>
  </METS:mdWrap>
</METS:dmdSec>
<METS:dmdSec ID="DM20">
  <METS:mdWrap MDTYPE="MODS">
    <METS:xmlData>
      <mods:mods>

<mods:titleInfo><mods:title>aboutme.htm_cmp_network010_bnr.gif</mods:title></mods:titleInfo>
      <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/aboutme.htm_cmp_network010_bnr.gif</mods:identifier>
      <mods:physicalDescription><mods:extent
type="filesize">1945</mods:extent></mods:physicalDescription>
      <mods:originInfo><mods:dateCaptured>2003-04-17T17:18:00</mods:dateCaptured></mods:originInfo>
      </mods:mods>
    </METS:xmlData>
  </METS:mdWrap>
</METS:dmdSec>
<METS:dmdSec ID="DM21">
  <METS:mdWrap MDTYPE="MODS">
    <METS:xmlData>
      <mods:mods>

<mods:titleInfo><mods:title>home_cmp_network010_vbtn.gif</mods:title></mods:titleInfo>
      <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/home_cmp_network010_vbtn.gif</mods:identifier>

```

```

                <mods:physicalDescription><mods:extent
type="filesize">166</mods:extent></mods:physicalDescription>
                <mods:originInfo><mods:dateCaptured>2003-04-
17T17:18:00</mods:dateCaptured></mods:originInfo>
                </mods:mods>
            </METS:xmlData>
        </METS:mdWrap>
    </METS:dmdSec>
    <METS:dmdSec ID="DM22">
        <METS:mdWrap MDTYPE="MODS">
            <METS:xmlData>
                <mods:mods>

<mods:titleInfo><mods:title>aboutme.htm_cmp_network010_vbtn_p.gif</mods:title></mods:titleInf
o>
                <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/aboutme.htm_cmp_
network010_vbtn_p.gif</mods:identifier>
                <mods:physicalDescription><mods:extent
type="filesize">247</mods:extent></mods:physicalDescription>
                <mods:originInfo><mods:dateCaptured>2003-04-
17T17:18:00</mods:dateCaptured></mods:originInfo>
                </mods:mods>
            </METS:xmlData>
        </METS:mdWrap>
    </METS:dmdSec>
    <METS:dmdSec ID="DM23">
        <METS:mdWrap MDTYPE="MODS">
            <METS:xmlData>
                <mods:mods>

<mods:titleInfo><mods:title>agenda.htm_cmp_network010_bnr.gif</mods:title></mods:titleInfo>
                <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/agenda.htm_cmp_n
etwork010_bnr.gif</mods:identifier>
                <mods:physicalDescription><mods:extent
type="filesize">2177</mods:extent></mods:physicalDescription>
                <mods:originInfo><mods:dateCaptured>2003-04-
17T17:18:00</mods:dateCaptured></mods:originInfo>
                </mods:mods>
            </METS:xmlData>
        </METS:mdWrap>
    </METS:dmdSec>
    <METS:dmdSec ID="DM24">
        <METS:mdWrap MDTYPE="MODS">
            <METS:xmlData>
                <mods:mods>

<mods:titleInfo><mods:title>agenda.htm_cmp_network010_vbtn_p.gif</mods:title></mods:titleInfo
>
                <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/agenda.htm_cmp_n
etwork010_vbtn_p.gif</mods:identifier>
                <mods:physicalDescription><mods:extent
type="filesize">298</mods:extent></mods:physicalDescription>
                <mods:originInfo><mods:dateCaptured>2003-04-
17T17:18:00</mods:dateCaptured></mods:originInfo>
                </mods:mods>
            </METS:xmlData>
        </METS:mdWrap>
    </METS:dmdSec>
    <METS:dmdSec ID="DM25">
        <METS:mdWrap MDTYPE="MODS">

```

```
<METS:xmlData>
  <mods:mods>

<mods:titleInfo><mods:title>favorite.htm_cmp_network010_bnr.gif</mods:title></mods:titleInfo>
  <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/favorite.htm_cmp
_network010_bnr.gif</mods:identifier>
  <mods:physicalDescription><mods:extent
type="filesize">1839</mods:extent></mods:physicalDescription>
  <mods:originInfo><mods:dateCaptured>2003-04-
17T17:18:00</mods:dateCaptured></mods:originInfo>
  </mods:mods>
</METS:xmlData>
</METS:mdWrap>
</METS:dmdSec>
<METS:dmdSec ID="DM26">
  <METS:mdWrap MDTYPE="MODS">
    <METS:xmlData>
      <mods:mods>

<mods:titleInfo><mods:title>favorite.htm_cmp_network010_vbtn_p.gif</mods:title></mods:titleIn
fo>
  <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/favorite.htm_cmp
_network010_vbtn_p.gif</mods:identifier>
  <mods:physicalDescription><mods:extent
type="filesize">235</mods:extent></mods:physicalDescription>
  <mods:originInfo><mods:dateCaptured>2003-04-
17T17:18:00</mods:dateCaptured></mods:originInfo>
  </mods:mods>
</METS:xmlData>
</METS:mdWrap>
</METS:dmdSec>
<METS:dmdSec ID="DM27">
  <METS:mdWrap MDTYPE="MODS">
    <METS:xmlData>
      <mods:mods>

<mods:titleInfo><mods:title>photo.htm_cmp_network010_bnr.gif</mods:title></mods:titleInfo>
  <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/photo.htm_cmp_ne
twork010_bnr.gif</mods:identifier>
  <mods:physicalDescription><mods:extent
type="filesize">1995</mods:extent></mods:physicalDescription>
  <mods:originInfo><mods:dateCaptured>2003-04-
17T17:18:00</mods:dateCaptured></mods:originInfo>
  </mods:mods>
</METS:xmlData>
</METS:mdWrap>
</METS:dmdSec>
<METS:dmdSec ID="DM28">
  <METS:mdWrap MDTYPE="MODS">
    <METS:xmlData>
      <mods:mods>

<mods:titleInfo><mods:title>photo.htm_cmp_network010_vbtn_p.gif</mods:title></mods:titleInfo>
  <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/photo.htm_cmp_ne
twork010_vbtn_p.gif</mods:identifier>
  <mods:physicalDescription><mods:extent
type="filesize">264</mods:extent></mods:physicalDescription>
  <mods:originInfo><mods:dateCaptured>2003-04-
17T17:18:00</mods:dateCaptured></mods:originInfo>
```

```

        </mods:mods>
      </METS:xmlData>
    </METS:mdWrap>
  </METS:dmdSec>
  <METS:dmdSec ID="DM29">
    <METS:mdWrap MDTYPE="MODS">
      <METS:xmlData>
        <mods:mods>

<mods:titleInfo><mods:title>feedback.htm_cmp_network010_bnr.gif</mods:title></mods:titleInfo>
      <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/feedback.htm_cmp_network010_bnr.gif</mods:identifier>
      <mods:physicalDescription><mods:extent
type="filesize">1854</mods:extent></mods:physicalDescription>
      <mods:originInfo><mods:dateCaptured>2003-04-17T17:18:00</mods:dateCaptured></mods:originInfo>
      </mods:mods>
    </METS:xmlData>
  </METS:mdWrap>
</METS:dmdSec>
  <METS:dmdSec ID="DM30">
    <METS:mdWrap MDTYPE="MODS">
      <METS:xmlData>
        <mods:mods>

<mods:titleInfo><mods:title>feedback.htm_cmp_network010_vbtn_p.gif</mods:title></mods:titleInfo>
      <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/feedback.htm_cmp_network010_vbtn_p.gif</mods:identifier>
      <mods:physicalDescription><mods:extent
type="filesize">237</mods:extent></mods:physicalDescription>
      <mods:originInfo><mods:dateCaptured>2003-04-17T17:18:00</mods:dateCaptured></mods:originInfo>
      </mods:mods>
    </METS:xmlData>
  </METS:mdWrap>
</METS:dmdSec>
  <METS:dmdSec ID="DM31">
    <METS:mdWrap MDTYPE="MODS">
      <METS:xmlData>
        <mods:mods>

<mods:titleInfo><mods:title>loretta_face.jpg</mods:title></mods:titleInfo>
      <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/loretta_face.jpg</mods:identifier>
      <mods:physicalDescription><mods:extent
type="filesize">8317</mods:extent></mods:physicalDescription>
      <mods:originInfo><mods:dateCaptured>2003-04-17T17:17:56</mods:dateCaptured></mods:originInfo>
      </mods:mods>
    </METS:xmlData>
  </METS:mdWrap>
</METS:dmdSec>
  <METS:dmdSec ID="DM32">
    <METS:mdWrap MDTYPE="MODS">
      <METS:xmlData>
        <mods:mods>
          <mods:titleInfo><mods:title>email.gif</mods:title></mods:titleInfo>

```



```

                <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/email.gif</mods:ident
ifier>
                <mods:physicalDescription><mods:extent
type="filesize">132</mods:extent></mods:physicalDescription>
                <mods:originInfo><mods:dateCaptured>2002-11-
25T11:01:32</mods:dateCaptured></mods:originInfo>
                </mods:mods>
            </METS:xmlData>
        </METS:mdWrap>
    </METS:dmdSec>
    <METS:dmdSec ID="DM33">
        <METS:mdWrap MDTYPE="MODS">
            <METS:xmlData>
                <mods:mods>
                    <mods:titleInfo><mods:title>new.gif</mods:title></mods:titleInfo>
                    <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/new.gif</mods:iden
tifier>
                    <mods:physicalDescription><mods:extent
type="filesize">3972</mods:extent></mods:physicalDescription>
                    <mods:originInfo><mods:dateCaptured>2002-11-
25T11:01:32</mods:dateCaptured></mods:originInfo>
                    </mods:mods>
                </METS:xmlData>
            </METS:mdWrap>
        </METS:dmdSec>
        <METS:dmdSec ID="DM34">
            <METS:mdWrap MDTYPE="MODS">
                <METS:xmlData>
                    <mods:mods>

<mods:titleInfo><mods:title>people_dallas.jpg</mods:title></mods:titleInfo>
                    <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/people_dallas.jpg<
/mods:identifier>
                    <mods:physicalDescription><mods:extent
type="filesize">189010</mods:extent></mods:physicalDescription>
                    <mods:originInfo><mods:dateCaptured>2003-04-
17T17:17:58</mods:dateCaptured></mods:originInfo>
                    </mods:mods>
                </METS:xmlData>
            </METS:mdWrap>
        </METS:dmdSec>
        <METS:dmdSec ID="DM35">
            <METS:mdWrap MDTYPE="MODS">
                <METS:xmlData>
                    <mods:mods>

<mods:titleInfo><mods:title>loretta_agenda.jpg</mods:title></mods:titleInfo>
                    <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/loretta_agenda.jpg
</mods:identifier>
                    <mods:physicalDescription><mods:extent
type="filesize">91669</mods:extent></mods:physicalDescription>
                    <mods:originInfo><mods:dateCaptured>2003-04-
17T17:17:52</mods:dateCaptured></mods:originInfo>
                    </mods:mods>
                </METS:xmlData>
            </METS:mdWrap>
        </METS:dmdSec>
        <METS:dmdSec ID="DM36">
            <METS:mdWrap MDTYPE="MODS">

```

```

    <METS:xmlData>
      <mods:mods>

<mods:titleInfo><mods:title>loretta_dallas.jpg</mods:title></mods:titleInfo>
      <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/loretta_dallas.jpg
</mods:identifier>
      <mods:physicalDescription><mods:extent
type="filesize">100152</mods:extent></mods:physicalDescription>
      <mods:originInfo><mods:dateCaptured>2003-04-
17T17:17:54</mods:dateCaptured></mods:originInfo>
      </mods:mods>
    </METS:xmlData>
  </METS:mdWrap>
</METS:dmdSec>
<METS:dmdSec ID="DM37">
  <METS:mdWrap MDTYPE="MODS">
    <METS:xmlData>
      <mods:mods>

<mods:titleInfo><mods:title>loretta_speak.jpg</mods:title></mods:titleInfo>
      <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/loretta_speak.jpg<
/mods:identifier>
      <mods:physicalDescription><mods:extent
type="filesize">119614</mods:extent></mods:physicalDescription>
      <mods:originInfo><mods:dateCaptured>2003-04-
17T17:17:56</mods:dateCaptured></mods:originInfo>
      </mods:mods>
    </METS:xmlData>
  </METS:mdWrap>
</METS:dmdSec>
<METS:dmdSec ID="DM38">
  <METS:mdWrap MDTYPE="MODS">
    <METS:xmlData>
      <mods:mods>

<mods:titleInfo><mods:title>loretta_mic2.jpg</mods:title></mods:titleInfo>
      <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/loretta_mic2.jpg</
mods:identifier>
      <mods:physicalDescription><mods:extent
type="filesize">155635</mods:extent></mods:physicalDescription>
      <mods:originInfo><mods:dateCaptured>2003-04-
17T17:17:56</mods:dateCaptured></mods:originInfo>
      </mods:mods>
    </METS:xmlData>
  </METS:mdWrap>
</METS:dmdSec>
<METS:dmdSec ID="DM39">
  <METS:mdWrap MDTYPE="MODS">
    <METS:xmlData>
      <mods:mods>

<mods:titleInfo><mods:title>loretta_bart.jpg</mods:title></mods:titleInfo>
      <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/loretta_bart.jpg</
mods:identifier>
      <mods:physicalDescription><mods:extent
type="filesize">232255</mods:extent></mods:physicalDescription>
      <mods:originInfo><mods:dateCaptured>2003-04-
17T17:17:54</mods:dateCaptured></mods:originInfo>
      </mods:mods>

```

```

        </METS:xmlData>
    </METS:mdWrap>
</METS:dmdSec>
<METS:dmdSec ID="DM40">
    <METS:mdWrap MDTYPE="MODS">
        <METS:xmlData>
            <mods:mods>
                <mods:titleInfo><mods:title>netbkgnd.gif</mods:title></mods:titleInfo>
                <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_themes/network/netbkgnd.
gif</mods:identifier>
                <mods:physicalDescription><mods:extent
type="filesize">1089</mods:extent></mods:physicalDescription>
                <mods:originInfo><mods:dateCaptured>2003-04-
17T17:17:50</mods:dateCaptured></mods:originInfo>
                </mods:mods>
            </METS:xmlData>
        </METS:mdWrap>
    </METS:dmdSec>
<METS:dmdSec ID="DM41">
    <METS:mdWrap MDTYPE="MODS">
        <METS:xmlData>
            <mods:mods>
                <mods:titleInfo><mods:title>anetbull.gif</mods:title></mods:titleInfo>
                <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_themes/network/anetbull.
gif</mods:identifier>
                <mods:physicalDescription><mods:extent
type="filesize">69</mods:extent></mods:physicalDescription>
                <mods:originInfo><mods:dateCaptured>2003-04-
17T17:17:50</mods:dateCaptured></mods:originInfo>
                </mods:mods>
            </METS:xmlData>
        </METS:mdWrap>
    </METS:dmdSec>
<METS:dmdSec ID="DM42">
    <METS:mdWrap MDTYPE="MODS">
        <METS:xmlData>
            <mods:mods>
                <mods:titleInfo><mods:title>anetbul2.gif</mods:title></mods:titleInfo>
                <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_themes/network/anetbul2.
gif</mods:identifier>
                <mods:physicalDescription><mods:extent
type="filesize">68</mods:extent></mods:physicalDescription>
                <mods:originInfo><mods:dateCaptured>2003-04-
17T17:17:50</mods:dateCaptured></mods:originInfo>
                </mods:mods>
            </METS:xmlData>
        </METS:mdWrap>
    </METS:dmdSec>
<METS:dmdSec ID="DM43">
    <METS:mdWrap MDTYPE="MODS">
        <METS:xmlData>
            <mods:mods>
                <mods:titleInfo><mods:title>filelist.xml</mods:title></mods:titleInfo>
                <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/index_files/filelist.xml<
/mods:identifier>
                <mods:physicalDescription><mods:extent
type="filesize">1955</mods:extent></mods:physicalDescription>
                <mods:originInfo><mods:dateCaptured>2003-03-
21T18:58:14</mods:dateCaptured></mods:originInfo>

```

```

        </mods:mods>
      </METS:xmlData>
    </METS:mdWrap>
  </METS:dmdSec>
  <METS:dmdSec ID="DM44">
    <METS:mdWrap MDTYPE="MODS">
      <METS:xmlData>
        <mods:mods>
          <mods:titleInfo><mods:title>image001.gif</mods:title></mods:titleInfo>
          <mods:identifier
type="uri">http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/index_files/image001.gif<
/mods:identifier>
          <mods:physicalDescription><mods:extent
type="filesize">5834</mods:extent></mods:physicalDescription>
          <mods:originInfo><mods:dateCaptured>2002-09-
04T14:12:28</mods:dateCaptured></mods:originInfo>
        </mods:mods>
      </METS:xmlData>
    </METS:mdWrap>
  </METS:dmdSec>
  <METS:amdSec>
    <METS:techMD ID="ADM1">
      <METS:mdWrap MDTYPE="OTHER" OTHERMDTYPE="textMD">
        <METS:xmlData>
          <textmd:textMD>
            <textmd:language>eng</textmd:language>
            <textmd:markup_language>HTML</textmd:markup_language>
            <textmd:textNote>index.html</textmd:textNote>
          </textmd:textMD>
        </METS:xmlData>
      </METS:mdWrap>
    </METS:techMD>
    <METS:techMD ID="ADM2">
      <METS:mdWrap MDTYPE="OTHER" OTHERMDTYPE="textMD">
        <METS:xmlData>
          <textmd:textMD>
            <textmd:language>eng</textmd:language>
            <textmd:markup_language>HTML</textmd:markup_language>
            <textmd:textNote>aboutme.htm</textmd:textNote>
          </textmd:textMD>
        </METS:xmlData>
      </METS:mdWrap>
    </METS:techMD>
    <METS:techMD ID="ADM3">
      <METS:mdWrap MDTYPE="OTHER" OTHERMDTYPE="textMD">
        <METS:xmlData>
          <textmd:textMD>
            <textmd:language>eng</textmd:language>
            <textmd:markup_language>HTML</textmd:markup_language>
            <textmd:textNote>agenda.htm</textmd:textNote>
          </textmd:textMD>
        </METS:xmlData>
      </METS:mdWrap>
    </METS:techMD>
    <METS:techMD ID="ADM4">
      <METS:mdWrap MDTYPE="OTHER" OTHERMDTYPE="textMD">
        <METS:xmlData>
          <textmd:textMD>
            <textmd:language>eng</textmd:language>
            <textmd:markup_language>HTML</textmd:markup_language>
            <textmd:textNote>favorite.htm</textmd:textNote>
          </textmd:textMD>
        </METS:xmlData>
      </METS:mdWrap>
    </METS:techMD>
  </METS:amdSec>
</METS:archive>

```

```

    </METS:mdWrap>
  </METS:techMD>
  <METS:techMD ID="ADM5">
    <METS:mdWrap MDTYPE="OTHER" OTHERMDTYPE="textMD">
      <METS:xmlData>
        <textmd:textMD>
          <textmd:language>eng</textmd:language>
          <textmd:markup_language>HTML</textmd:markup_language>
          <textmd:textNote>photo.htm</textmd:textNote>
        </textmd:textMD>
      </METS:xmlData>
    </METS:mdWrap>
  </METS:techMD>
  <METS:techMD ID="ADM6">
    <METS:mdWrap MDTYPE="OTHER" OTHERMDTYPE="textMD">
      <METS:xmlData>
        <textmd:textMD>
          <textmd:language>eng</textmd:language>
          <textmd:markup_language>HTML</textmd:markup_language>
          <textmd:textNote>feedback.htm</textmd:textNote>
        </textmd:textMD>
      </METS:xmlData>
    </METS:mdWrap>
  </METS:techMD>
  <METS:techMD ID="ADM7">
    <METS:mdWrap MDTYPE="OTHER" OTHERMDTYPE="textMD">
      <METS:xmlData>
        <textmd:textMD>
          <textmd:language>eng</textmd:language>
          <textmd:markup_language>HTML</textmd:markup_language>
          <textmd:textNote>index4062.html</textmd:textNote>
        </textmd:textMD>
      </METS:xmlData>
    </METS:mdWrap>
  </METS:techMD>
  <METS:techMD ID="ADM8">
    <METS:mdWrap MDTYPE="NISOIMG">
      <METS:xmlData>
        <mix:mix>
          <mix:BasicImageParameters>
            <mix:Format>
              <mix:MIMETYPE>image/gif</mix:MIMETYPE>
              <mix:Compression>
                <mix:CompressionScheme>1</mix:CompressionScheme>
                <mix:CompressionLevel>10</mix:CompressionLevel>
              </mix:Compression>
              <mix:PhotometricInterpretation>
                <mix:ColorSpace>2</mix:ColorSpace>
              </mix:PhotometricInterpretation>
            </mix:Format>
            <mix:File>
<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/home_c
mp_network010_vbtn_p.gif</mix:ImageIdentifier>
              <mix:FileSize>212</mix:FileSize>
            </mix:File>
            <mix:PreferredPresentation/>
          </mix:BasicImageParameters>
          <mix:ImageCreation/>
          <mix:ImagingPerformanceAssessment>
            <mix:SpatialMetrics>
              <mix:ImageWidth>140</mix:ImageWidth>
              <mix:ImageLength>24</mix:ImageLength>

```

```

        </mix:SpatialMetrics>
        <mix:Energetics>
            <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
        </mix:Energetics>
    </mix:ImagingPerformanceAssessment>
    <mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM9">
    <METS:mdWrap MDTYPE="NISOIMG">
        <METS:xmlData>
            <mix:mix>
                <mix:BasicImageParameters>
                    <mix:Format>
                        <mix:MIMEType>image/gif</mix:MIMEType>
                        <mix:Compression>
                            <mix:CompressionScheme>1</mix:CompressionScheme>
                            <mix:CompressionLevel>10</mix:CompressionLevel>
                        </mix:Compression>
                        <mix:PhotometricInterpretation>
                            <mix:ColorSpace>2</mix:ColorSpace>
                        </mix:PhotometricInterpretation>
                    </mix:Format>
                    <mix:File>

<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/home_c
mp_network010_vbtn_a.gif</mix:ImageIdentifier>
                        <mix:FileSize>218</mix:FileSize>
                    </mix:File>
                    <mix:PreferredPresentation/>
                </mix:BasicImageParameters>
                <mix:ImageCreation/>
                <mix:ImagingPerformanceAssessment>
                    <mix:SpatialMetrics>
                        <mix:ImageWidth>140</mix:ImageWidth>
                        <mix:ImageLength>24</mix:ImageLength>
                    </mix:SpatialMetrics>
                    <mix:Energetics>
                        <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
                    </mix:Energetics>
                </mix:ImagingPerformanceAssessment>
                <mix:ChangeHistory/>
            </mix:mix>
        </METS:xmlData>
    </METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM10">
    <METS:mdWrap MDTYPE="NISOIMG">
        <METS:xmlData>
            <mix:mix>
                <mix:BasicImageParameters>
                    <mix:Format>
                        <mix:MIMEType>image/gif</mix:MIMEType>
                        <mix:Compression>
                            <mix:CompressionScheme>1</mix:CompressionScheme>
                            <mix:CompressionLevel>10</mix:CompressionLevel>
                        </mix:Compression>
                        <mix:PhotometricInterpretation>
                            <mix:ColorSpace>2</mix:ColorSpace>
                        </mix:PhotometricInterpretation>
                    </mix:Format>

```

```

        <mix:File>

<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/aboutm
e.htm_cmp_network010_vbtn.gif</mix:ImageIdentifier>
        <mix:FileSize>204</mix:FileSize>
        </mix:File>
        <mix:PreferredPresentation/>
</mix:BasicImageParameters>
<mix:ImageCreation/>
<mix:ImagingPerformanceAssessment>
        <mix:SpatialMetrics>
                <mix:ImageWidth>140</mix:ImageWidth>
                <mix:ImageLength>24</mix:ImageLength>
        </mix:SpatialMetrics>
        <mix:Energetics>
                <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
        </mix:Energetics>
</mix:ImagingPerformanceAssessment>
        <mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM11">
        <METS:mdWrap MDTYPE="NISOIMG">
                <METS:xmlData>
                        <mix:mix>
                                <mix:BasicImageParameters>
                                        <mix:Format>
                                                <mix:MIMETYPE>image/gif</mix:MIMETYPE>
                                                <mix:Compression>
                                                        <mix:CompressionScheme>1</mix:CompressionScheme>
                                                        <mix:CompressionLevel>10</mix:CompressionLevel>
                                                </mix:Compression>
                                                <mix:PhotometricInterpretation>
                                                        <mix:ColorSpace>2</mix:ColorSpace>
                                                </mix:PhotometricInterpretation>
                                        </mix:Format>
                                <mix:File>

<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/aboutm
e.htm_cmp_network010_vbtn_a.gif</mix:ImageIdentifier>
                                <mix:FileSize>251</mix:FileSize>
                                </mix:File>
                                <mix:PreferredPresentation/>
</mix:BasicImageParameters>
<mix:ImageCreation/>
<mix:ImagingPerformanceAssessment>
                                <mix:SpatialMetrics>
                                        <mix:ImageWidth>140</mix:ImageWidth>
                                        <mix:ImageLength>24</mix:ImageLength>
                                </mix:SpatialMetrics>
                                <mix:Energetics>
                                        <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
                                </mix:Energetics>
</mix:ImagingPerformanceAssessment>
                                <mix:ChangeHistory/>
                        </mix:mix>
                </METS:xmlData>
        </METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM12">
        <METS:mdWrap MDTYPE="NISOIMG">

```

```

<METS:xmlData>
  <mix:mix>
    <mix:BasicImageParameters>
      <mix:Format>
        <mix:MIMETYPE>image/gif</mix:MIMETYPE>
        <mix:Compression>
          <mix:CompressionScheme>1</mix:CompressionScheme>
          <mix:CompressionLevel>10</mix:CompressionLevel>
        </mix:Compression>
        <mix:PhotometricInterpretation>
          <mix:ColorSpace>2</mix:ColorSpace>
        </mix:PhotometricInterpretation>
      </mix:Format>
    </mix:mix>
  </METS:xmlData>
  <mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/agenda.htm_cmp_network010_vbtn.gif</mix:ImageIdentifier>
  <mix:FileSize>248</mix:FileSize>
  </mix:File>
  <mix:PreferredPresentation/>
</mix:BasicImageParameters>
<mix:ImageCreation/>
<mix:ImagingPerformanceAssessment>
  <mix:SpatialMetrics>
    <mix:ImageWidth>140</mix:ImageWidth>
    <mix:ImageLength>24</mix:ImageLength>
  </mix:SpatialMetrics>
  <mix:Energetics>
    <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
  </mix:Energetics>
</mix:ImagingPerformanceAssessment>
<mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM13">
  <METS:mdWrap MDTYPE="NISOIMG">
    <METS:xmlData>
      <mix:mix>
        <mix:BasicImageParameters>
          <mix:Format>
            <mix:MIMETYPE>image/gif</mix:MIMETYPE>
            <mix:Compression>
              <mix:CompressionScheme>1</mix:CompressionScheme>
              <mix:CompressionLevel>10</mix:CompressionLevel>
            </mix:Compression>
            <mix:PhotometricInterpretation>
              <mix:ColorSpace>2</mix:ColorSpace>
            </mix:PhotometricInterpretation>
          </mix:Format>
        </mix:mix>
      </METS:xmlData>
      <mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/agenda.htm_cmp_network010_vbtn_a.gif</mix:ImageIdentifier>
      <mix:FileSize>303</mix:FileSize>
      </mix:File>
      <mix:PreferredPresentation/>
    </mix:BasicImageParameters>
    <mix:ImageCreation/>
    <mix:ImagingPerformanceAssessment>
      <mix:SpatialMetrics>
        <mix:ImageWidth>140</mix:ImageWidth>

```



```

        <mix:ImageLength>24</mix:ImageLength>
    </mix:SpatialMetrics>
    <mix:Energetics>
        <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
    </mix:Energetics>
</mix:ImagingPerformanceAssessment>
    <mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM14">
    <METS:mdWrap MDTYPE="NISOIMG">
        <METS:xmlData>
            <mix:mix>
                <mix:BasicImageParameters>
                    <mix:Format>
                        <mix:MIMEType>image/gif</mix:MIMEType>
                        <mix:Compression>
                            <mix:CompressionScheme>1</mix:CompressionScheme>
                            <mix:CompressionLevel>10</mix:CompressionLevel>
                        </mix:Compression>
                        <mix:PhotometricInterpretation>
                            <mix:ColorSpace>2</mix:ColorSpace>
                        </mix:PhotometricInterpretation>
                    </mix:Format>
                    <mix:File>
<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/favori
te.htm_cmp_network010_vbtn.gif</mix:ImageIdentifier>
                        <mix:FileSize>191</mix:FileSize>
                    </mix:File>
                    <mix:PreferredPresentation/>
                </mix:BasicImageParameters>
                <mix:ImageCreation/>
                <mix:ImagingPerformanceAssessment>
                    <mix:SpatialMetrics>
                        <mix:ImageWidth>140</mix:ImageWidth>
                        <mix:ImageLength>24</mix:ImageLength>
                    </mix:SpatialMetrics>
                    <mix:Energetics>
                        <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
                    </mix:Energetics>
                </mix:ImagingPerformanceAssessment>
                <mix:ChangeHistory/>
            </mix:mix>
        </METS:xmlData>
    </METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM15">
    <METS:mdWrap MDTYPE="NISOIMG">
        <METS:xmlData>
            <mix:mix>
                <mix:BasicImageParameters>
                    <mix:Format>
                        <mix:MIMEType>image/gif</mix:MIMEType>
                        <mix:Compression>
                            <mix:CompressionScheme>1</mix:CompressionScheme>
                            <mix:CompressionLevel>10</mix:CompressionLevel>
                        </mix:Compression>
                        <mix:PhotometricInterpretation>
                            <mix:ColorSpace>2</mix:ColorSpace>
                        </mix:PhotometricInterpretation>

```

```

        </mix:Format>
        <mix:File>

<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/favori
te.htm_cmp_network010_vbtn_a.gif</mix:ImageIdentifier>
        <mix:FileSize>239</mix:FileSize>
        </mix:File>
        <mix:PreferredPresentation/>
</mix:BasicImageParameters>
<mix:ImageCreation/>
<mix:ImagingPerformanceAssessment>
        <mix:SpatialMetrics>
                <mix:ImageWidth>140</mix:ImageWidth>
                <mix:ImageLength>24</mix:ImageLength>
        </mix:SpatialMetrics>
        <mix:Energetics>
                <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
        </mix:Energetics>
</mix:ImagingPerformanceAssessment>
<mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM16">
        <METS:mdWrap MDTYPE="NISOIMG">
                <METS:xmlData>
                        <mix:mix>
                                <mix:BasicImageParameters>
                                        <mix:Format>
                                                <mix:MIMETYPE>image/gif</mix:MIMETYPE>
                                                <mix:Compression>
                                                        <mix:CompressionScheme>1</mix:CompressionScheme>
                                                        <mix:CompressionLevel>10</mix:CompressionLevel>
                                                </mix:Compression>
                                                <mix:PhotometricInterpretation>
                                                        <mix:ColorSpace>2</mix:ColorSpace>
                                                </mix:PhotometricInterpretation>
                                        </mix:Format>
                                        <mix:File>
                                                <mix:FileSize>216</mix:FileSize>
                                        </mix:File>
                                        <mix:PreferredPresentation/>
                                </mix:BasicImageParameters>
                                <mix:ImageCreation/>
                                <mix:ImagingPerformanceAssessment>
                                        <mix:SpatialMetrics>
                                                <mix:ImageWidth>140</mix:ImageWidth>
                                                <mix:ImageLength>24</mix:ImageLength>
                                        </mix:SpatialMetrics>
                                        <mix:Energetics>
                                                <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
                                        </mix:Energetics>
                                </mix:ImagingPerformanceAssessment>
                                <mix:ChangeHistory/>
                        </mix:mix>
                </METS:xmlData>
        </METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM17">
        <METS:mdWrap MDTYPE="NISOIMG">
                <METS:xmlData>
                        <mix:mix>

```

```

    <mix:BasicImageParameters>
      <mix:Format>
        <mix:MIMEType>image/gif</mix:MIMEType>
        <mix:Compression>
          <mix:CompressionScheme>1</mix:CompressionScheme>
          <mix:CompressionLevel>10</mix:CompressionLevel>
        </mix:Compression>
        <mix:PhotometricInterpretation>
          <mix:ColorSpace>2</mix:ColorSpace>
        </mix:PhotometricInterpretation>
      </mix:Format>
      <mix:File>
        <mix:FileSize>268</mix:FileSize>
      </mix:File>
      <mix:PreferredPresentation/>
    </mix:BasicImageParameters>
    <mix:ImageCreation/>
    <mix:ImagingPerformanceAssessment>
      <mix:SpatialMetrics>
        <mix:ImageWidth>140</mix:ImageWidth>
        <mix:ImageLength>24</mix:ImageLength>
      </mix:SpatialMetrics>
      <mix:Energetics>
        <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
      </mix:Energetics>
    </mix:ImagingPerformanceAssessment>
    <mix:ChangeHistory/>
  </mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM18">
  <METS:mdWrap MDTYPE="NISOIMG">
    <METS:xmlData>
      <mix:mix>
        <mix:BasicImageParameters>
          <mix:Format>
            <mix:MIMEType>image/gif</mix:MIMEType>
            <mix:Compression>
              <mix:CompressionScheme>1</mix:CompressionScheme>
              <mix:CompressionLevel>10</mix:CompressionLevel>
            </mix:Compression>
            <mix:PhotometricInterpretation>
              <mix:ColorSpace>2</mix:ColorSpace>
            </mix:PhotometricInterpretation>
          </mix:Format>
          <mix:File>
            <mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/\_derived/feedback.htm\_cmp\_network010\_vbtn.gif</mix:ImageIdentifier>
            <mix:FileSize>194</mix:FileSize>
          </mix:File>
          <mix:PreferredPresentation/>
        </mix:BasicImageParameters>
        <mix:ImageCreation/>
        <mix:ImagingPerformanceAssessment>
          <mix:SpatialMetrics>
            <mix:ImageWidth>140</mix:ImageWidth>
            <mix:ImageLength>24</mix:ImageLength>
          </mix:SpatialMetrics>
          <mix:Energetics>
            <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
          </mix:Energetics>
        </mix:ImagingPerformanceAssessment>
      </mix:mix>
    </METS:xmlData>
  </METS:mdWrap>
</METS:techMD>

```

```

        </mix:ImagingPerformanceAssessment>
        <mix:ChangeHistory/>
    </mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM19">
    <METS:mdWrap MDTYPE="NISOIMG">
        <METS:xmlData>
            <mix:mix>
                <mix:BasicImageParameters>
                    <mix:Format>
                        <mix:MIMETYPE>image/gif</mix:MIMETYPE>
                        <mix:Compression>
                            <mix:CompressionScheme>1</mix:CompressionScheme>
                            <mix:CompressionLevel>10</mix:CompressionLevel>
                        </mix:Compression>
                        <mix:PhotometricInterpretation>
                            <mix:ColorSpace>2</mix:ColorSpace>
                        </mix:PhotometricInterpretation>
                    </mix:Format>
                    <mix:File>

<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/feedba
ck.htm_cmp_network010_vbtn_a.gif</mix:ImageIdentifier>
                        <mix:FileSize>240</mix:FileSize>
                    </mix:File>
                    <mix:PreferredPresentation/>
                </mix:BasicImageParameters>
                <mix:ImageCreation/>
                <mix:ImagingPerformanceAssessment>
                    <mix:SpatialMetrics>
                        <mix:ImageWidth>140</mix:ImageWidth>
                        <mix:ImageLength>24</mix:ImageLength>
                    </mix:SpatialMetrics>
                    <mix:Energetics>
                        <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
                    </mix:Energetics>
                </mix:ImagingPerformanceAssessment>
                <mix:ChangeHistory/>
            </mix:mix>
        </METS:xmlData>
    </METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM20">
    <METS:mdWrap MDTYPE="NISOIMG">
        <METS:xmlData>
            <mix:mix>
                <mix:BasicImageParameters>
                    <mix:Format>
                        <mix:MIMETYPE>image/gif</mix:MIMETYPE>
                        <mix:Compression>
                            <mix:CompressionScheme>1</mix:CompressionScheme>
                            <mix:CompressionLevel>10</mix:CompressionLevel>
                        </mix:Compression>
                        <mix:PhotometricInterpretation>
                            <mix:ColorSpace>2</mix:ColorSpace>
                        </mix:PhotometricInterpretation>
                    </mix:Format>
                    <mix:File>

<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/aboutm
e.htm_cmp_network010_bnr.gif</mix:ImageIdentifier>
                </mix:File>
            </mix:mix>
        </METS:xmlData>
    </METS:mdWrap>
</METS:techMD>

```

```

        <mix:FileSize>1945</mix:FileSize>
    </mix:File>
    <mix:PreferredPresentation/>
</mix:BasicImageParameters>
<mix:ImageCreation/>
<mix:ImagingPerformanceAssessment>
    <mix:SpatialMetrics>
        <mix:ImageWidth>598</mix:ImageWidth>
        <mix:ImageLength>141</mix:ImageLength>
    </mix:SpatialMetrics>
    <mix:Energetics>
        <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
    </mix:Energetics>
</mix:ImagingPerformanceAssessment>
    <mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM21">
    <METS:mdWrap MDTYPE="NISOIMG">
        <METS:xmlData>
            <mix:mix>
                <mix:BasicImageParameters>
                    <mix:Format>
                        <mix:MIMETYPE>image/gif</mix:MIMETYPE>
                        <mix:Compression>
                            <mix:CompressionScheme>1</mix:CompressionScheme>
                            <mix:CompressionLevel>10</mix:CompressionLevel>
                        </mix:Compression>
                        <mix:PhotometricInterpretation>
                            <mix:ColorSpace>2</mix:ColorSpace>
                        </mix:PhotometricInterpretation>
                    </mix:Format>
                    <mix:File>
<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/home_c
mp_network010_vbtn.gif</mix:ImageIdentifier>
                        <mix:FileSize>166</mix:FileSize>
                    </mix:File>
                    <mix:PreferredPresentation/>
                </mix:BasicImageParameters>
                <mix:ImageCreation/>
                <mix:ImagingPerformanceAssessment>
                    <mix:SpatialMetrics>
                        <mix:ImageWidth>140</mix:ImageWidth>
                        <mix:ImageLength>24</mix:ImageLength>
                    </mix:SpatialMetrics>
                    <mix:Energetics>
                        <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
                    </mix:Energetics>
                </mix:ImagingPerformanceAssessment>
                <mix:ChangeHistory/>
            </mix:mix>
        </METS:xmlData>
    </METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM22">
    <METS:mdWrap MDTYPE="NISOIMG">
        <METS:xmlData>
            <mix:mix>
                <mix:BasicImageParameters>
                    <mix:Format>

```

```

        <mix:MIMEType>image/gif</mix:MIMEType>
        <mix:Compression>
            <mix:CompressionScheme>1</mix:CompressionScheme>
            <mix:CompressionLevel>10</mix:CompressionLevel>
        </mix:Compression>
        <mix:PhotometricInterpretation>
            <mix:ColorSpace>2</mix:ColorSpace>
        </mix:PhotometricInterpretation>
    </mix:Format>
</mix:File>

<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/aboutm
e.htm_cmp_network010_vbtn_p.gif</mix:ImageIdentifier>
    <mix:FileSize>247</mix:FileSize>
</mix:File>
    <mix:PreferredPresentation/>
</mix:BasicImageParameters>
<mix:ImageCreation/>
<mix:ImagingPerformanceAssessment>
    <mix:SpatialMetrics>
        <mix:ImageWidth>140</mix:ImageWidth>
        <mix:ImageLength>24</mix:ImageLength>
    </mix:SpatialMetrics>
    <mix:Energetics>
        <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
    </mix:Energetics>
</mix:ImagingPerformanceAssessment>
    <mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM23">
    <METS:mdWrap MDTYPE="NISOIMG">
        <METS:xmlData>
            <mix:mix>
                <mix:BasicImageParameters>
                    <mix:Format>
                        <mix:MIMEType>image/gif</mix:MIMEType>
                        <mix:Compression>
                            <mix:CompressionScheme>1</mix:CompressionScheme>
                            <mix:CompressionLevel>10</mix:CompressionLevel>
                        </mix:Compression>
                        <mix:PhotometricInterpretation>
                            <mix:ColorSpace>2</mix:ColorSpace>
                        </mix:PhotometricInterpretation>
                    </mix:Format>
                    <mix:File>

<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/agenda
.htm_cmp_network010_bnr.gif</mix:ImageIdentifier>
    <mix:FileSize>2177</mix:FileSize>
</mix:File>
    <mix:PreferredPresentation/>
</mix:BasicImageParameters>
<mix:ImageCreation/>
<mix:ImagingPerformanceAssessment>
    <mix:SpatialMetrics>
        <mix:ImageWidth>598</mix:ImageWidth>
        <mix:ImageLength>141</mix:ImageLength>
    </mix:SpatialMetrics>
    <mix:Energetics>
        <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
    </mix:Energetics>
</mix:ImagingPerformanceAssessment>
    <mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
</METS:techMD ID="ADM23">
    <METS:mdWrap MDTYPE="NISOIMG">
        <METS:xmlData>
            <mix:mix>
                <mix:BasicImageParameters>
                    <mix:Format>
                        <mix:MIMEType>image/gif</mix:MIMEType>
                        <mix:Compression>
                            <mix:CompressionScheme>1</mix:CompressionScheme>
                            <mix:CompressionLevel>10</mix:CompressionLevel>
                        </mix:Compression>
                        <mix:PhotometricInterpretation>
                            <mix:ColorSpace>2</mix:ColorSpace>
                        </mix:PhotometricInterpretation>
                    </mix:Format>
                    <mix:File>

```

```

        </mix:Energetics>
    </mix:ImagingPerformanceAssessment>
    <mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM24">
    <METS:mdWrap MDTYPE="NISOIMG">
        <METS:xmlData>
            <mix:mix>
                <mix:BasicImageParameters>
                    <mix:Format>
                        <mix:MIMETYPE>image/gif</mix:MIMETYPE>
                        <mix:Compression>
                            <mix:CompressionScheme>1</mix:CompressionScheme>
                            <mix:CompressionLevel>10</mix:CompressionLevel>
                        </mix:Compression>
                        <mix:PhotometricInterpretation>
                            <mix:ColorSpace>2</mix:ColorSpace>
                        </mix:PhotometricInterpretation>
                    </mix:Format>
                    <mix:File>
<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/agenda
.htm_cmp_network010_vbtn_p.gif</mix:ImageIdentifier>
                <mix:FileSize>298</mix:FileSize>
            </mix:File>
            <mix:PreferredPresentation/>
        </mix:BasicImageParameters>
        <mix:ImageCreation/>
        <mix:ImagingPerformanceAssessment>
            <mix:SpatialMetrics>
                <mix:ImageWidth>140</mix:ImageWidth>
                <mix:ImageLength>24</mix:ImageLength>
            </mix:SpatialMetrics>
            <mix:Energetics>
                <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
            </mix:Energetics>
        </mix:ImagingPerformanceAssessment>
        <mix:ChangeHistory/>
    </mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM25">
    <METS:mdWrap MDTYPE="NISOIMG">
        <METS:xmlData>
            <mix:mix>
                <mix:BasicImageParameters>
                    <mix:Format>
                        <mix:MIMETYPE>image/gif</mix:MIMETYPE>
                        <mix:Compression>
                            <mix:CompressionScheme>1</mix:CompressionScheme>
                            <mix:CompressionLevel>10</mix:CompressionLevel>
                        </mix:Compression>
                        <mix:PhotometricInterpretation>
                            <mix:ColorSpace>2</mix:ColorSpace>
                        </mix:PhotometricInterpretation>
                    </mix:Format>
                    <mix:File>

```

```

<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/favori
te.htm_cmp_network010_bnr.gif</mix:ImageIdentifier>
  <mix:FileSize>1839</mix:FileSize>
  </mix:File>
  <mix:PreferredPresentation/>
</mix:BasicImageParameters>
<mix:ImageCreation/>
<mix:ImagingPerformanceAssessment>
  <mix:SpatialMetrics>
    <mix:ImageWidth>598</mix:ImageWidth>
    <mix:ImageLength>141</mix:ImageLength>
  </mix:SpatialMetrics>
  <mix:Energetics>
    <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
  </mix:Energetics>
</mix:ImagingPerformanceAssessment>
  <mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM26">
  <METS:mdWrap MDTYPE="NISOIMG">
    <METS:xmlData>
      <mix:mix>
        <mix:BasicImageParameters>
          <mix:Format>
            <mix:MIMEType>image/gif</mix:MIMEType>
            <mix:Compression>
              <mix:CompressionScheme>1</mix:CompressionScheme>
              <mix:CompressionLevel>10</mix:CompressionLevel>
            </mix:Compression>
            <mix:PhotometricInterpretation>
              <mix:ColorSpace>2</mix:ColorSpace>
            </mix:PhotometricInterpretation>
          </mix:Format>
          <mix:File>

<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/favori
te.htm_cmp_network010_vbtn_p.gif</mix:ImageIdentifier>
  <mix:FileSize>235</mix:FileSize>
  </mix:File>
  <mix:PreferredPresentation/>
</mix:BasicImageParameters>
<mix:ImageCreation/>
<mix:ImagingPerformanceAssessment>
  <mix:SpatialMetrics>
    <mix:ImageWidth>140</mix:ImageWidth>
    <mix:ImageLength>24</mix:ImageLength>
  </mix:SpatialMetrics>
  <mix:Energetics>
    <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
  </mix:Energetics>
</mix:ImagingPerformanceAssessment>
  <mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM27">
  <METS:mdWrap MDTYPE="NISOIMG">
    <METS:xmlData>

```



```

    <mix:mix>
      <mix:BasicImageParameters>
        <mix:Format>
          <mix:MIMEType>image/gif</mix:MIMEType>
          <mix:Compression>
            <mix:CompressionScheme>1</mix:CompressionScheme>
            <mix:CompressionLevel>10</mix:CompressionLevel>
          </mix:Compression>
          <mix:PhotometricInterpretation>
            <mix:ColorSpace>2</mix:ColorSpace>
          </mix:PhotometricInterpretation>
        </mix:Format>
      <mix:File>

<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/photo.
htm_cmp_network010_bnr.gif</mix:ImageIdentifier>
      <mix:FileSize>1995</mix:FileSize>
    </mix:File>
    <mix:PreferredPresentation/>
  </mix:BasicImageParameters>
  <mix:ImageCreation/>
  <mix:ImagingPerformanceAssessment>
    <mix:SpatialMetrics>
      <mix:ImageWidth>598</mix:ImageWidth>
      <mix:ImageLength>141</mix:ImageLength>
    </mix:SpatialMetrics>
    <mix:Energetics>
      <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
    </mix:Energetics>
  </mix:ImagingPerformanceAssessment>
  <mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM28">
  <METS:mdWrap MDTYPE="NISOIMG">
    <METS:xmlData>
      <mix:mix>
        <mix:BasicImageParameters>
          <mix:Format>
            <mix:MIMEType>image/gif</mix:MIMEType>
            <mix:Compression>
              <mix:CompressionScheme>1</mix:CompressionScheme>
              <mix:CompressionLevel>10</mix:CompressionLevel>
            </mix:Compression>
            <mix:PhotometricInterpretation>
              <mix:ColorSpace>2</mix:ColorSpace>
            </mix:PhotometricInterpretation>
          </mix:Format>
        <mix:File>

<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/photo.
htm_cmp_network010_vbtn_p.gif</mix:ImageIdentifier>
      <mix:FileSize>264</mix:FileSize>
    </mix:File>
    <mix:PreferredPresentation/>
  </mix:BasicImageParameters>
  <mix:ImageCreation/>
  <mix:ImagingPerformanceAssessment>
    <mix:SpatialMetrics>
      <mix:ImageWidth>140</mix:ImageWidth>
      <mix:ImageLength>24</mix:ImageLength>

```

```

        </mix:SpatialMetrics>
        <mix:Energetics>
            <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
        </mix:Energetics>
    </mix:ImagingPerformanceAssessment>
    <mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM29">
    <METS:mdWrap MDTYPE="NISOIMG">
        <METS:xmlData>
            <mix:mix>
                <mix:BasicImageParameters>
                    <mix:Format>
                        <mix:MIMETYPE>image/gif</mix:MIMETYPE>
                        <mix:Compression>
                            <mix:CompressionScheme>1</mix:CompressionScheme>
                            <mix:CompressionLevel>10</mix:CompressionLevel>
                        </mix:Compression>
                        <mix:PhotometricInterpretation>
                            <mix:ColorSpace>2</mix:ColorSpace>
                        </mix:PhotometricInterpretation>
                    </mix:Format>
                    <mix:File>

<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/feedba
ck.htm_cmp_network010_bnr.gif</mix:ImageIdentifier>
                        <mix:FileSize>1854</mix:FileSize>
                    </mix:File>
                    <mix:PreferredPresentation/>
                </mix:BasicImageParameters>
                <mix:ImageCreation/>
                <mix:ImagingPerformanceAssessment>
                    <mix:SpatialMetrics>
                        <mix:ImageWidth>598</mix:ImageWidth>
                        <mix:ImageLength>141</mix:ImageLength>
                    </mix:SpatialMetrics>
                </mix:ImagingPerformanceAssessment>
                <mix:ChangeHistory/>
            </mix:mix>
        </METS:xmlData>
    </METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM30">
    <METS:mdWrap MDTYPE="NISOIMG">
        <METS:xmlData>
            <mix:mix>
                <mix:BasicImageParameters>
                    <mix:Format>
                        <mix:MIMETYPE>image/gif</mix:MIMETYPE>
                        <mix:Compression>
                            <mix:CompressionScheme>1</mix:CompressionScheme>
                            <mix:CompressionLevel>10</mix:CompressionLevel>
                        </mix:Compression>
                        <mix:PhotometricInterpretation>
                            <mix:ColorSpace>2</mix:ColorSpace>
                        </mix:PhotometricInterpretation>
                    </mix:Format>
                    <mix:File>

```

```

<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/feedba
ck.htm_cmp_network010_vbtn_p.gif</mix:ImageIdentifier>
  <mix:FileSize>237</mix:FileSize>
  </mix:File>
  <mix:PreferredPresentation/>
</mix:BasicImageParameters>
<mix:ImageCreation/>
<mix:ImagingPerformanceAssessment>
  <mix:SpatialMetrics>
    <mix:ImageWidth>140</mix:ImageWidth>
    <mix:ImageLength>24</mix:ImageLength>
  </mix:SpatialMetrics>
  <mix:Energetics>
    <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
  </mix:Energetics>
</mix:ImagingPerformanceAssessment>
  <mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM31">
  <METS:mdWrap MDTYPE="NISOIMG">
    <METS:xmlData>
      <mix:mix>
        <mix:BasicImageParameters>
          <mix:Format>
            <mix:MIMEType>image/jpeg</mix:MIMEType>
            <mix:Compression>
              <mix:CompressionScheme>1</mix:CompressionScheme>
              <mix:CompressionLevel>10</mix:CompressionLevel>
            </mix:Compression>
            <mix:PhotometricInterpretation>
              <mix:ColorSpace>2</mix:ColorSpace>
            </mix:PhotometricInterpretation>
          </mix:Format>
          <mix:File>

<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/loretta_
face.jpg</mix:ImageIdentifier>
  <mix:FileSize>8317</mix:FileSize>
  </mix:File>
  <mix:PreferredPresentation/>
</mix:BasicImageParameters>
<mix:ImageCreation/>
<mix:ImagingPerformanceAssessment>
  <mix:SpatialMetrics>
    <mix:ImageWidth>143</mix:ImageWidth>
    <mix:ImageLength>186</mix:ImageLength>
  </mix:SpatialMetrics>
  <mix:Energetics>
    <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
  </mix:Energetics>
</mix:ImagingPerformanceAssessment>
  <mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM32">
  <METS:mdWrap MDTYPE="NISOIMG">
    <METS:xmlData>

```

```

<mix:mix>
  <mix:BasicImageParameters>
    <mix:Format>
      <mix:MIMETYPE>image/gif</mix:MIMETYPE>
      <mix:Compression>
        <mix:CompressionScheme>1</mix:CompressionScheme>
        <mix:CompressionLevel>10</mix:CompressionLevel>
      </mix:Compression>
      <mix:PhotometricInterpretation>
        <mix:ColorSpace>2</mix:ColorSpace>
      </mix:PhotometricInterpretation>
    </mix:Format>
  </mix:BasicImageParameters>
  <mix:File>
    <mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/email.gif</mix:ImageIdentifier>
    <mix:FileSize>132</mix:FileSize>
  </mix:File>
  <mix:PreferredPresentation/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM33">
  <METS:mdWrap MDTYPE="NISOIMG">
    <METS:xmlData>
      <mix:mix>
        <mix:BasicImageParameters>
          <mix:Format>
            <mix:MIMETYPE>image/gif</mix:MIMETYPE>
            <mix:Compression>
              <mix:CompressionScheme>1</mix:CompressionScheme>
              <mix:CompressionLevel>10</mix:CompressionLevel>
            </mix:Compression>
            <mix:PhotometricInterpretation>
              <mix:ColorSpace>2</mix:ColorSpace>
            </mix:PhotometricInterpretation>
          </mix:Format>
        </mix:BasicImageParameters>
        <mix:File>
          <mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/new.gif</mix:ImageIdentifier>
          <mix:FileSize>3972</mix:FileSize>
        </mix:File>
        <mix:PreferredPresentation/>
      </mix:mix>
    </METS:xmlData>
  </METS:mdWrap>
</METS:techMD>

```

```

        </mix:SpatialMetrics>
        <mix:Energetics>
            <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
        </mix:Energetics>
    </mix:ImagingPerformanceAssessment>
    <mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM34">
    <METS:mdWrap MDTYPE="NISOIMG">
        <METS:xmlData>
            <mix:mix>
                <mix:BasicImageParameters>
                    <mix:Format>
                        <mix:MIMETYPE>image/jpeg</mix:MIMETYPE>
                        <mix:Compression>
                            <mix:CompressionScheme>1</mix:CompressionScheme>
                            <mix:CompressionLevel>10</mix:CompressionLevel>
                        </mix:Compression>
                        <mix:PhotometricInterpretation>
                            <mix:ColorSpace>2</mix:ColorSpace>
                        </mix:PhotometricInterpretation>
                    </mix:Format>
                    <mix:File>
<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/people_d
allas.jpg</mix:ImageIdentifier>
                        <mix:FileSize>189010</mix:FileSize>
                    </mix:File>
                    <mix:PreferredPresentation/>
                </mix:BasicImageParameters>
                <mix:ImageCreation/>
                <mix:ImagingPerformanceAssessment>
                    <mix:SpatialMetrics>
                        <mix:ImageWidth>1020</mix:ImageWidth>
                        <mix:ImageLength>767</mix:ImageLength>
                    </mix:SpatialMetrics>
                    <mix:Energetics>
                        <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
                    </mix:Energetics>
                </mix:ImagingPerformanceAssessment>
                <mix:ChangeHistory/>
            </mix:mix>
        </METS:xmlData>
    </METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM35">
    <METS:mdWrap MDTYPE="NISOIMG">
        <METS:xmlData>
            <mix:mix>
                <mix:BasicImageParameters>
                    <mix:Format>
                        <mix:MIMETYPE>image/jpeg</mix:MIMETYPE>
                        <mix:Compression>
                            <mix:CompressionScheme>1</mix:CompressionScheme>
                            <mix:CompressionLevel>10</mix:CompressionLevel>
                        </mix:Compression>
                        <mix:PhotometricInterpretation>
                            <mix:ColorSpace>2</mix:ColorSpace>
                        </mix:PhotometricInterpretation>
                    </mix:Format>

```

```

        <mix:File>
<mix:ImageIdentifier>:/www.aniagolu.org/images/loretta_agenda.jpg</mix:ImageIdentifier>
        <mix:FileSize>91669</mix:FileSize>
        </mix:File>
        <mix:PreferredPresentation/>
</mix:BasicImageParameters>
<mix:ImageCreation/>
<mix:ImagingPerformanceAssessment>
        <mix:SpatialMetrics>
                <mix:ImageWidth>633</mix:ImageWidth>
                <mix:ImageLength>692</mix:ImageLength>
        </mix:SpatialMetrics>
        <mix:Energetics>
                <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
        </mix:Energetics>
</mix:ImagingPerformanceAssessment>
        <mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM36">
        <METS:mdWrap MDTYPE="NISOIMG">
                <METS:xmlData>
                        <mix:mix>
                                <mix:BasicImageParameters>
                                        <mix:Format>
                                                <mix:MIMEType>image/jpeg</mix:MIMEType>
                                                <mix:Compression>
                                                        <mix:CompressionScheme>1</mix:CompressionScheme>
                                                        <mix:CompressionLevel>10</mix:CompressionLevel>
                                                </mix:Compression>
                                                <mix:PhotometricInterpretation>
                                                        <mix:ColorSpace>2</mix:ColorSpace>
                                                </mix:PhotometricInterpretation>
                                        </mix:Format>
                                        <mix:File>
<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/loretta_
dallas.jpg</mix:ImageIdentifier>
                                <mix:FileSize>100152</mix:FileSize>
                                </mix:File>
                                <mix:PreferredPresentation/>
</mix:BasicImageParameters>
<mix:ImageCreation/>
<mix:ImagingPerformanceAssessment>
        <mix:SpatialMetrics>
                <mix:ImageWidth>557</mix:ImageWidth>
                <mix:ImageLength>853</mix:ImageLength>
        </mix:SpatialMetrics>
        <mix:Energetics>
                <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
        </mix:Energetics>
</mix:ImagingPerformanceAssessment>
        <mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM37">
        <METS:mdWrap MDTYPE="NISOIMG">
                <METS:xmlData>

```

```

<mix:mix>
  <mix:BasicImageParameters>
    <mix:Format>
      <mix:MIMETYPE>image/jpeg</mix:MIMETYPE>
      <mix:Compression>
        <mix:CompressionScheme>1</mix:CompressionScheme>
        <mix:CompressionLevel>10</mix:CompressionLevel>
      </mix:Compression>
      <mix:PhotometricInterpretation>
        <mix:ColorSpace>2</mix:ColorSpace>
      </mix:PhotometricInterpretation>
    </mix:Format>
  <mix:File>

<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/loretta_
speak.jpg</mix:ImageIdentifier>
  <mix:FileSize>119614</mix:FileSize>
  </mix:File>
  <mix:PreferredPresentation/>
</mix:BasicImageParameters>
<mix:ImageCreation/>
  <mix:ImagingPerformanceAssessment>
    <mix:SpatialMetrics>
      <mix:ImageWidth>677</mix:ImageWidth>
      <mix:ImageLength>877</mix:ImageLength>
    </mix:SpatialMetrics>
    <mix:Energetics>
      <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
    </mix:Energetics>
  </mix:ImagingPerformanceAssessment>
  <mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM38">
  <METS:mdWrap MDTYPE="NISOIMG">
    <METS:xmlData>
      <mix:mix>
        <mix:BasicImageParameters>
          <mix:Format>
            <mix:MIMETYPE>image/jpeg</mix:MIMETYPE>
            <mix:Compression>
              <mix:CompressionScheme>1</mix:CompressionScheme>
              <mix:CompressionLevel>10</mix:CompressionLevel>
            </mix:Compression>
            <mix:PhotometricInterpretation>
              <mix:ColorSpace>2</mix:ColorSpace>
            </mix:PhotometricInterpretation>
          </mix:Format>
        <mix:File>

<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/loretta_
mic2.jpg</mix:ImageIdentifier>
  <mix:FileSize>155635</mix:FileSize>
  </mix:File>
  <mix:PreferredPresentation/>
</mix:BasicImageParameters>
<mix:ImageCreation/>
  <mix:ImagingPerformanceAssessment>
    <mix:SpatialMetrics>
      <mix:ImageWidth>829</mix:ImageWidth>
      <mix:ImageLength>1057</mix:ImageLength>
    </mix:SpatialMetrics>
  </mix:ImagingPerformanceAssessment>
  <mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
</METS:MDList>
</METS:MDWrap>

```

```

        </mix:SpatialMetrics>
        <mix:Energetics>
            <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
        </mix:Energetics>
    </mix:ImagingPerformanceAssessment>
    <mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM39">
    <METS:mdWrap MDTYPE="NISOIMG">
        <METS:xmlData>
            <mix:mix>
                <mix:BasicImageParameters>
                    <mix:Format>
                        <mix:MIMEType>image/jpeg</mix:MIMEType>
                        <mix:Compression>
                            <mix:CompressionScheme>1</mix:CompressionScheme>
                            <mix:CompressionLevel>10</mix:CompressionLevel>
                        </mix:Compression>
                        <mix:PhotometricInterpretation>
                            <mix:ColorSpace>2</mix:ColorSpace>
                        </mix:PhotometricInterpretation>
                    </mix:Format>
                    <mix:File>
<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/loretta_
bart.jpg</mix:ImageIdentifier>
                        <mix:FileSize>232255</mix:FileSize>
                    </mix:File>
                    <mix:PreferredPresentation/>
                </mix:BasicImageParameters>
                <mix:ImageCreation/>
                <mix:ImagingPerformanceAssessment>
                    <mix:SpatialMetrics>
                        <mix:ImageWidth>1207</mix:ImageWidth>
                        <mix:ImageLength>991</mix:ImageLength>
                    </mix:SpatialMetrics>
                    <mix:Energetics>
                        <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
                    </mix:Energetics>
                </mix:ImagingPerformanceAssessment>
                <mix:ChangeHistory/>
            </mix:mix>
        </METS:xmlData>
    </METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM40">
    <METS:mdWrap MDTYPE="NISOIMG">
        <METS:xmlData>
            <mix:mix>
                <mix:BasicImageParameters>
                    <mix:Format>
                        <mix:MIMEType>image/gif</mix:MIMEType>
                        <mix:Compression>
                            <mix:CompressionScheme>1</mix:CompressionScheme>
                            <mix:CompressionLevel>10</mix:CompressionLevel>
                        </mix:Compression>
                        <mix:PhotometricInterpretation>
                            <mix:ColorSpace>2</mix:ColorSpace>
                        </mix:PhotometricInterpretation>
                    </mix:Format>

```



```

        <mix:File>
<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_themes/network
/netbkgnd.gif</mix:ImageIdentifier>
        <mix:FileSize>1089</mix:FileSize>
        </mix:File>
        <mix:PreferredPresentation/>
</mix:BasicImageParameters>
<mix:ImageCreation/>
<mix:ImagingPerformanceAssessment>
        <mix:SpatialMetrics>
                <mix:ImageWidth>1600</mix:ImageWidth>
                <mix:ImageLength>5</mix:ImageLength>
        </mix:SpatialMetrics>
        <mix:Energetics>
                <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
        </mix:Energetics>
</mix:ImagingPerformanceAssessment>
        <mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM41">
        <METS:mdWrap MDTYPE="NISOIMG">
                <METS:xmlData>
                        <mix:mix>
                                <mix:BasicImageParameters>
                                        <mix:Format>
                                                <mix:MIMEMimeType>image/gif</mix:MIMEMimeType>
                                                <mix:Compression>
                                                        <mix:CompressionScheme>1</mix:CompressionScheme>
                                                        <mix:CompressionLevel>10</mix:CompressionLevel>
                                                </mix:Compression>
                                                <mix:PhotometricInterpretation>
                                                        <mix:ColorSpace>2</mix:ColorSpace>
                                                </mix:PhotometricInterpretation>
                                        </mix:Format>
                                <mix:File>
<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_themes/network
/anetbull.gif</mix:ImageIdentifier>
                                <mix:FileSize>69</mix:FileSize>
                                </mix:File>
                                <mix:PreferredPresentation/>
</mix:BasicImageParameters>
<mix:ImageCreation/>
<mix:ImagingPerformanceAssessment>
                                <mix:SpatialMetrics>
                                        <mix:ImageWidth>12</mix:ImageWidth>
                                        <mix:ImageLength>12</mix:ImageLength>
                                </mix:SpatialMetrics>
                                <mix:Energetics>
                                        <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
                                </mix:Energetics>
</mix:ImagingPerformanceAssessment>
                                <mix:ChangeHistory/>
                        </mix:mix>
                </METS:xmlData>
        </METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM42">
        <METS:mdWrap MDTYPE="NISOIMG">

```

```

<METS:xmlData>
  <mix:mix>
    <mix:BasicImageParameters>
      <mix:Format>
        <mix:MIMETYPE>image/gif</mix:MIMETYPE>
        <mix:Compression>
          <mix:CompressionScheme>1</mix:CompressionScheme>
          <mix:CompressionLevel>10</mix:CompressionLevel>
        </mix:Compression>
        <mix:PhotometricInterpretation>
          <mix:ColorSpace>2</mix:ColorSpace>
        </mix:PhotometricInterpretation>
      </mix:Format>
    </mix:File>

<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_themes/network
/anetbul2.gif</mix:ImageIdentifier>
    <mix:FileSize>68</mix:FileSize>
  </mix:File>
  <mix:PreferredPresentation/>
</mix:BasicImageParameters>
<mix:ImageCreation/>
<mix:ImagingPerformanceAssessment>
  <mix:SpatialMetrics>
    <mix:ImageWidth>12</mix:ImageWidth>
    <mix:ImageLength>12</mix:ImageLength>
  </mix:SpatialMetrics>
  <mix:Energetics>
    <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
  </mix:Energetics>
</mix:ImagingPerformanceAssessment>
  <mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM43">
  <METS:mdWrap MDTYPE="NISOIMG">
    <METS:xmlData>
      <textmd:textMD>
        <textmd:language>eng</textmd:language>
        <textmd:markup_language>XML</textmd:markup_language>
        <textmd:textNote>filelist.xml</textmd:textNote>
      </textmd:textMD>
    </METS:xmlData>
  </METS:mdWrap>
</METS:techMD>
<METS:techMD ID="ADM44">
  <METS:mdWrap MDTYPE="NISOIMG">
    <METS:xmlData>
      <mix:mix>
        <mix:BasicImageParameters>
          <mix:Format>
            <mix:MIMETYPE>image/gif</mix:MIMETYPE>
            <mix:Compression>
              <mix:CompressionScheme>1</mix:CompressionScheme>
              <mix:CompressionLevel>10</mix:CompressionLevel>
            </mix:Compression>
            <mix:PhotometricInterpretation>
              <mix:ColorSpace>2</mix:ColorSpace>
            </mix:PhotometricInterpretation>
          </mix:Format>
        </mix:File>

```

```

<mix:ImageIdentifier>http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/index_files/ima
ge001.gif</mix:ImageIdentifier>
    <mix:FileSize>5834</mix:FileSize>
  </mix:File>
  <mix:PreferredPresentation/>
</mix:BasicImageParameters>
<mix:ImageCreation/>
<mix:ImagingPerformanceAssessment>
  <mix:SpatialMetrics>
    <mix:ImageWidth>336</mix:ImageWidth>
    <mix:ImageLength>110</mix:ImageLength>
  </mix:SpatialMetrics>
  <mix:Energetics>
    <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
  </mix:Energetics>
</mix:ImagingPerformanceAssessment>
<mix:ChangeHistory/>
</mix:mix>
</METS:xmlData>
</METS:mdWrap>
</METS:techMD>
</METS:amdSec>
<METS:fileSec>
  <METS:fileGrp>
    <METS:file ADMID="ADM1" CREATED="2003-09-09T18:58:16" ID="FID1"
MIMETYPE="text/html">
      <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/index.html"/>
    </METS:file>
    <METS:file ADMID="ADM2" CREATED="2002-09-09T14:32:26" ID="FID2"
MIMETYPE="text/html">
      <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/aboutme.htm"/>
    </METS:file>
    <METS:file ADMID="ADM3" CREATED="2002-09-09T14:32:26" ID="FID3"
MIMETYPE="text/html">
      <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/agenda.htm"/>
    </METS:file>
    <METS:file ADMID="ADM4" CREATED="2002-09-09T14:32:26" ID="FID4"
MIMETYPE="text/html">
      <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/favorite.htm"/>
    </METS:file>
    <METS:file ADMID="ADM5" CREATED="2002-09-09T14:32:26" ID="FID5"
MIMETYPE="text/html">
      <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/photo.htm"/>
    </METS:file>
    <METS:file ADMID="ADM6" CREATED="2002-09-09T14:32:26" ID="FID6"
MIMETYPE="text/html">
      <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/feedback.htm"/>
    </METS:file>
    <METS:file ADMID="ADM7" CREATED="2003-04-19T10:36:26" ID="FID7"
MIMETYPE="text/html">
      <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_vti_bin/fpcount.exe/ind
ex4062.html"/>
    </METS:file>
    <METS:file ADMID="ADM8" CREATED="2003-04-19T17:18:00" ID="FID8"
MIMETYPE="image/gif">

```

```

        <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/home_cmp_networ
k010_vbtn_p.gif"/>
        </METS:file>
        <METS:file ADMID="ADM9" CREATED="2003-04-19T17:18:00" ID="FID9"
MIMETYPE="image/gif">
        <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/home_cmp_networ
k010_vbtn_a.gif"/>
        </METS:file>
        <METS:file ADMID="ADM10" CREATED="2003-04-19T17:18:00" ID="FID10"
MIMETYPE="text/html">
        <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/aboutme.htm_cmp
_network010_vbtn.gif"/>
        </METS:file>
        <METS:file ADMID="ADM11" CREATED="2003-04-19T17:18:00" ID="FID11"
MIMETYPE="text/html">
        <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/aboutme.htm_cmp
_network010_vbtn_a.gif"/>
        </METS:file>
        <METS:file ADMID="ADM12" CREATED="2003-04-19T17:18:00" ID="FID12"
MIMETYPE="text/html">
        <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/agenda.htm_cmp_
network010_vbtn.gif"/>
        </METS:file>
        <METS:file ADMID="ADM13" CREATED="2003-04-19T17:18:00" ID="FID13"
MIMETYPE="text/html">
        <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/agenda.htm_cmp_
network010_vbtn_a.gif"/>
        </METS:file>
        <METS:file ADMID="ADM14" CREATED="2003-04-19T17:18:00" ID="FID14"
MIMETYPE="text/html">
        <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/favorite.htm_cm
p_network010_vbtn.gif"/>
        </METS:file>
        <METS:file ADMID="ADM15" CREATED="2003-04-19T17:18:00" ID="FID15"
MIMETYPE="text/html">
        <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/favorite.htm_cm
p_network010_vbtn_a.gif"/>
        </METS:file>
        <METS:file ADMID="ADM16" CREATED="2003-04-19T17:18:00" ID="FID16"
MIMETYPE="text/html">
        <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/photo.htm_cmp_n
etwork010_vbtn.gif"/>
        </METS:file>
        <METS:file ADMID="ADM17" CREATED="2003-04-19T17:18:00" ID="FID17"
MIMETYPE="text/html">
        <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/photo.htm_cmp_n
etwork010_vbtn_a.gif"/>
        </METS:file>
        <METS:file ADMID="ADM18" CREATED="2003-04-19T17:18:00" ID="FID18"
MIMETYPE="text/html">
        <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/feedback.htm_cm
p_network010_vbtn.gif"/>
```

```
</METS:file>
  <METS:file ADMID="ADM19" CREATED="2003-04-19T17:18:00" ID="FID19"
MIMETYPE="text/html">
  <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/feedback.htm_cm
p_network010_vbtn_a.gif"/>
  </METS:file>
  <METS:file ADMID="ADM20" CREATED="2003-04-19T17:18:00" ID="FID20"
MIMETYPE="text/html">
  <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/aboutme.htm_cmp
_network010_bnr.gif"/>
  </METS:file>
  <METS:file ADMID="ADM21" CREATED="2003-04-19T17:18:00" ID="FID21"
MIMETYPE="image/gif">
  <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/home_cmp_networ
k010_vbtn.gif"/>
  </METS:file>
  <METS:file ADMID="ADM22" CREATED="2003-04-19T17:18:00" ID="FID22"
MIMETYPE="text/html">
  <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/aboutme.htm_cmp
_network010_vbtn_p.gif"/>
  </METS:file>
  <METS:file ADMID="ADM23" CREATED="2003-04-19T17:18:00" ID="FID23"
MIMETYPE="text/html">
  <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/agenda.htm_cmp_
network010_bnr.gif"/>
  </METS:file>
  <METS:file ADMID="ADM24" CREATED="2003-04-19T17:18:00" ID="FID24"
MIMETYPE="text/html">
  <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/agenda.htm_cmp_
network010_vbtn_p.gif"/>
  </METS:file>
  <METS:file ADMID="ADM25" CREATED="2003-04-19T17:18:00" ID="FID25"
MIMETYPE="text/html">
  <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/favorite.htm_cm
p_network010_bnr.gif"/>
  </METS:file>
  <METS:file ADMID="ADM26" CREATED="2003-04-19T17:18:00" ID="FID26"
MIMETYPE="text/html">
  <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/favorite.htm_cm
p_network010_vbtn_p.gif"/>
  </METS:file>
  <METS:file ADMID="ADM27" CREATED="2003-04-19T17:18:00" ID="FID27"
MIMETYPE="text/html">
  <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/photo.htm_cmp_n
etwork010_bnr.gif"/>
  </METS:file>
  <METS:file ADMID="ADM28" CREATED="2003-04-19T17:18:00" ID="FID28"
MIMETYPE="text/html">
  <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/photo.htm_cmp_n
etwork010_vbtn_p.gif"/>
  </METS:file>
  <METS:file ADMID="ADM29" CREATED="2003-04-19T17:18:00" ID="FID29"
MIMETYPE="text/html">
```

```
<METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/feedback.htm_cm
p_network010_bnr.gif"/>
</METS:file>
<METS:file ADMID="ADM30" CREATED="2003-04-19T17:18:00" ID="FID30"
MIMETYPE="text/html">
  <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_derived/feedback.htm_cm
p_network010_vbtn_p.gif"/>
  </METS:file>
  <METS:file ADMID="ADM31" CREATED="2003-04-19T17:17:56" ID="FID31"
MIMETYPE="image/jpeg">
    <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/loretta_face.jpg"
/>
    </METS:file>
    <METS:file ADMID="ADM32" CREATED="2002-11-25T11:01:32" ID="FID32"
MIMETYPE="image/gif">
      <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/email.gif"/>
      </METS:file>
      <METS:file ADMID="ADM33" CREATED="2002-11-25T11:01:32" ID="FID33"
MIMETYPE="image/gif">
        <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/new.gif"/>
        </METS:file>
        <METS:file ADMID="ADM34" CREATED="2003-04-19T17:17:58" ID="FID34"
MIMETYPE="image/jpeg">
          <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/people_dallas.jpg
"/>
          </METS:file>
          <METS:file ADMID="ADM35" CREATED="2003-04-19T17:17:52" ID="FID35"
MIMETYPE="image/jpeg">
            <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/loretta_agenda.jp
g"/>
            </METS:file>
            <METS:file ADMID="ADM36" CREATED="2003-04-19T17:17:54" ID="FID36"
MIMETYPE="image/jpeg">
              <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/loretta_dallas.jp
g"/>
              </METS:file>
              <METS:file ADMID="ADM37" CREATED="2003-04-19T17:17:56" ID="FID37"
MIMETYPE="image/jpeg">
                <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/loretta_speak.jpg
"/>
                </METS:file>
                <METS:file ADMID="ADM38" CREATED="2003-04-19T17:17:56" ID="FID38"
MIMETYPE="image/jpeg">
                  <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/loretta_mic2.jpg"
/>
                  </METS:file>
                  <METS:file ADMID="ADM39" CREATED="2003-04-19T17:17:54" ID="FID39"
MIMETYPE="image/jpeg">
                    <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/images/loretta_bart.jpg"
/>
                    </METS:file>
```

```

    <METS:file ADMID="ADM40" CREATED="2003-04-19T17:17:50" ID="FID40"
MIMETYPE="image/gif">
    <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_themes/network/netbkgnd
.gif"/>
    </METS:file>
    <METS:file ADMID="ADM41" CREATED="2003-04-19T17:17:50" ID="FID41"
MIMETYPE="image/gif">
    <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_themes/network/anetbull
.gif"/>
    </METS:file>
    <METS:file ADMID="ADM42" CREATED="2003-04-19T17:17:50" ID="FID42"
MIMETYPE="image/gif">
    <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/_themes/network/anetbul2
.gif"/>
    </METS:file>
    <METS:file ADMID="ADM43" CREATED="2003-04-19T18:58:14" ID="FID43" MIMETYPE="xml">
    <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/index_files/filelist.xml
"/>
    </METS:file>
    <METS:file ADMID="ADM44" CREATED="2002-04-19T14:12:28" ID="FID44"
MIMETYPE="image/gif">
    <METS:FLocat LOCTYPE="URL"
xlink:href="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/index_files/image001.gif
"/>
    </METS:file>
  </METS:fileGrp>
</METS:fileSec>
<METS:structMap>
  <METS:div DMDID="DM01 DM1" ID="PAGE1"
    LABEL="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/index.html">
    <METS:fptr>
      <!-- the page file along with its embedded image/graphic links below -->
      <METS:par>
        <METS:area FILEID="FID1" />
        <METS:area FILEID="FID8" />
        <METS:area FILEID="FID10" />
        <METS:area FILEID="FID12" />
        <METS:area FILEID="FID14" />
        <METS:area FILEID="FID16" />
        <METS:area FILEID="FID18" />
        <METS:area FILEID="FID31" />
        <METS:area FILEID="FID32" />
        <METS:area FILEID="FID33" />
        <METS:area FILEID="FID34" />
        <METS:area FILEID="FID40" />
        <METS:area FILEID="FID41" />
        <METS:area FILEID="FID42" />
      </METS:par>
    </METS:fptr>
    <!-- the area pointers should be to the actual byte offsets of the HREF etc in
the HTML file but here they are not (yet) -->
    <METS:div ID="LINK1" LABEL="Link to Who Is She">
      <METS:fptr>
        <METS:area BEGIN="100" BETYPE="BYTE" END="120" FILEID="FID1"/>
      </METS:fptr>
    </METS:div>
    <METS:div ID="LINK2" LABEL="Link to Agenda, Not Gender">
      <METS:fptr>
        <METS:area BEGIN="200" BETYPE="BYTE" END="218" FILEID="FID1"/>
      </METS:fptr>
    </METS:div>
  </METS:div>
</METS:structMap>

```

```

    </METS:fptr>
</METS:div>
<METS:div ID="LINK3" LABEL="Link to Favorites">
  <METS:fptr>
    <METS:area BEGIN="300" BETYPE="BYTE" END="309" FILEID="FID1"/>
  </METS:fptr>
</METS:div>
<METS:div ID="LINK4" LABEL="Link to Photo Gallery">
  <METS:fptr>
    <METS:area BEGIN="400" BETYPE="BYTE" END="413" FILEID="FID1"/>
  </METS:fptr>
</METS:div>
<METS:div ID="LINK5" LABEL="Link to Feedback">
  <METS:fptr>
    <METS:area BEGIN="500" BETYPE="BYTE" END="508" FILEID="FID1"/>
  </METS:fptr>
</METS:div>
<METS:div DMDID="DM2" ID="PAGE2"
  LABEL="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/aboutme.htm">
  <METS:fptr>
    <METS:par>
      <METS:area FILEID="FID2"/>
      <METS:area FILEID="FID20"/>
      <METS:area FILEID="FID22"/>
      <METS:area FILEID="FID12"/>
      <METS:area FILEID="FID14"/>
      <METS:area FILEID="FID18"/>
      <METS:area FILEID="FID21"/>
      <METS:area FILEID="FID16"/>
      <METS:area FILEID="FID41"/>
      <METS:area FILEID="FID40"/>
      <METS:area FILEID="FID31"/>
    </METS:par>
  </METS:fptr>
<METS:div ID="LINK6" LABEL="Home">
  <METS:fptr>
    <METS:area BEGIN="100" BETYPE="BYTE" END="120" FILEID="FID2"/>
  </METS:fptr>
</METS:div>
<METS:div ID="LINK7" LABEL="Link to Agenda, Not Gender">
  <METS:fptr>
    <METS:area BEGIN="200" BETYPE="BYTE" END="218" FILEID="FID2"/>
  </METS:fptr>
</METS:div>
<METS:div ID="LINK8" LABEL="Link to Favorites">
  <METS:fptr>
    <METS:area BEGIN="300" BETYPE="BYTE" END="309" FILEID="FID2"/>
  </METS:fptr>
</METS:div>
<METS:div ID="LINK9" LABEL="Link to Photo Gallery">
  <METS:fptr>
    <METS:area BEGIN="400" BETYPE="BYTE" END="413" FILEID="FID2"/>
  </METS:fptr>
</METS:div>
<METS:div ID="LINK10" LABEL="Link to Feedback">
  <METS:fptr>
    <METS:area BEGIN="500" BETYPE="BYTE" END="508" FILEID="FID2"/>
  </METS:fptr>
</METS:div>
</METS:div>
<METS:div DMDID="DM3" ID="PAGE3"
  LABEL="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/agenda.htm">
  <METS:fptr>

```



```

<METS:par>
  <METS:area FILEID="FID3" />
  <METS:area FILEID="FID8" />
  <METS:area FILEID="FID23" />
  <METS:area FILEID="FID24" />
  <METS:area FILEID="FID14" />
  <METS:area FILEID="FID18" />
  <METS:area FILEID="FID16" />
  <METS:area FILEID="FID21" />
  <METS:area FILEID="FID40" />
  <METS:area FILEID="FID41" />
  <METS:area FILEID="FID42" />
  <METS:area FILEID="FID35" />
</METS:par>
</METS:fptr>
<METS:div ID="LINK11" LABEL="Home">
  <METS:fptr>
    <METS:area BEGIN="100" BETYPE="BYTE" END="120" FILEID="FID2" />
  </METS:fptr>
</METS:div>
<METS:div ID="LINK12" LABEL="Link to Who Is She">
  <METS:fptr>
    <METS:area BEGIN="100" BETYPE="BYTE" END="120" FILEID="FID3" />
  </METS:fptr>
</METS:div>
<METS:div ID="LINK13" LABEL="Link to Favorites">
  <METS:fptr>
    <METS:area BEGIN="300" BETYPE="BYTE" END="309" FILEID="FID3" />
  </METS:fptr>
</METS:div>
<METS:div ID="LINK14" LABEL="Link to Photo Gallery">
  <METS:fptr>
    <METS:area BEGIN="400" BETYPE="BYTE" END="413" FILEID="FID3" />
  </METS:fptr>
</METS:div>
<METS:div ID="LINK15" LABEL="Link to Feedback">
  <METS:fptr>
    <METS:area BEGIN="500" BETYPE="BYTE" END="508" FILEID="FID3" />
  </METS:fptr>
</METS:div>
</METS:div>
<METS:div DMDID="DM4" ID="PAGE4"

```

LABEL="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/favorite.htm">

```

  <METS:fptr>
    <METS:par>
      <METS:area FILEID="FID4" />
      <METS:area FILEID="FID8" />
      <METS:area FILEID="FID23" />
      <METS:area FILEID="FID18" />
      <METS:area FILEID="FID16" />
      <METS:area FILEID="FID21" />
      <METS:area FILEID="FID25" />
      <METS:area FILEID="FID26" />
      <METS:area FILEID="FID40" />
      <METS:area FILEID="FID41" />
      <METS:area FILEID="FID42" />
      <METS:area FILEID="FID31" />
    </METS:par>
  </METS:fptr>
<METS:div ID="LINK16" LABEL="Home">
  <METS:fptr>
    <METS:area BEGIN="100" BETYPE="BYTE" END="120" FILEID="FID2" />

```

```

        </METS:fptr>
    </METS:div>
    <METS:div ID="LINK17" LABEL="Link to Who Is She">
        <METS:fptr>
            <METS:area BEGIN="100" BETYPE="BYTE" END="120" FILEID="FID4" />
        </METS:fptr>
    </METS:div>
    <METS:div ID="LINK18" LABEL="Link to Agenda, Not Gender">
        <METS:fptr>
            <METS:area BEGIN="200" BETYPE="BYTE" END="218" FILEID="FID4" />
        </METS:fptr>
    </METS:div>
    <METS:div ID="LINK19" LABEL="Link to Photo Gallery">
        <METS:fptr>
            <METS:area BEGIN="400" BETYPE="BYTE" END="413" FILEID="FID4" />
        </METS:fptr>
    </METS:div>
    <METS:div ID="LINK20" LABEL="Link to Feedback">
        <METS:fptr>
            <METS:area BEGIN="500" BETYPE="BYTE" END="508" FILEID="FID4" />
        </METS:fptr>
    </METS:div>
</METS:div>
<METS:div DMDID="DM5" ID="PAGE5"
    LABEL="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/photo.htm">
    <METS:fptr>
        <METS:par>
            <METS:area FILEID="FID5" />
            <METS:area FILEID="FID8" />
            <METS:area FILEID="FID23" />
            <METS:area FILEID="FID18" />
            <METS:area FILEID="FID16" />
            <METS:area FILEID="FID21" />
            <METS:area FILEID="FID27" />
            <METS:area FILEID="FID28" />
            <METS:area FILEID="FID42" />
            <METS:area FILEID="FID39" />
            <METS:area FILEID="FID36" />
            <METS:area FILEID="FID37" />
            <METS:area FILEID="FID38" />
            <METS:area FILEID="FID34" />
        </METS:par>
    </METS:fptr>
    <METS:div ID="LINK21" LABEL="Home">
        <METS:fptr>
            <METS:area BEGIN="100" BETYPE="BYTE" END="120" FILEID="FID5" />
        </METS:fptr>
    </METS:div>
    <METS:div ID="LINK22" LABEL="Link to Who Is She">
        <METS:fptr>
            <METS:area BEGIN="100" BETYPE="BYTE" END="120" FILEID="FID5" />
        </METS:fptr>
    </METS:div>
    <METS:div ID="LINK23" LABEL="Link to Agenda, Not Gender">
        <METS:fptr>
            <METS:area BEGIN="200" BETYPE="BYTE" END="218" FILEID="FID5" />
        </METS:fptr>
    </METS:div>
    <METS:div ID="LINK24" LABEL="Link to Favorites">
        <METS:fptr>
            <METS:area BEGIN="300" BETYPE="BYTE" END="309" FILEID="FID5" />
        </METS:fptr>
    </METS:div>

```

```

<METS:div ID="LINK25" LABEL="Link to Feedback">
  <METS:fptr>
    <METS:area BEGIN="500" BETYPE="BYTE" END="508" FILEID="FID5" />
  </METS:fptr>
</METS:div>
</METS:div>
<METS:div DMDID="DM6" ID="PAGE6"
  LABEL="http://dlib.nyu.edu/webarchive/metstest/www.aniagolu.org/feedback.htm">
  <METS:fptr>
    <METS:par>
      <METS:area FILEID="FID6" />
      <METS:area FILEID="FID8" />
      <METS:area FILEID="FID23" />
      <METS:area FILEID="FID25" />
      <METS:area FILEID="FID26" />
      <METS:area FILEID="FID18" />
      <METS:area FILEID="FID16" />
      <METS:area FILEID="FID21" />
      <METS:area FILEID="FID40" />
      <METS:area FILEID="FID41" />
      <METS:area FILEID="FID42" />
      <METS:area FILEID="FID37" />
    </METS:par>
  </METS:fptr>
  <METS:div ID="LINK26" LABEL="Home">
    <METS:fptr>
      <METS:area BEGIN="100" BETYPE="BYTE" END="120" FILEID="FID5" />
    </METS:fptr>
  </METS:div>
  <METS:div ID="LINK27" LABEL="Link to Who Is She">
    <METS:fptr>
      <METS:area BEGIN="100" BETYPE="BYTE" END="120" FILEID="FID6" />
    </METS:fptr>
  </METS:div>
  <METS:div ID="LINK28" LABEL="Link to Agenda, Not Gender">
    <METS:fptr>
      <METS:area BEGIN="200" BETYPE="BYTE" END="218" FILEID="FID6" />
    </METS:fptr>
  </METS:div>
  <METS:div ID="LINK29" LABEL="Link to Favorites">
    <METS:fptr>
      <METS:area BEGIN="300" BETYPE="BYTE" END="309" FILEID="FID6" />
    </METS:fptr>
  </METS:div>
  <METS:div ID="LINK30" LABEL="Link to Photo Gallery">
    <METS:fptr>
      <METS:area BEGIN="400" BETYPE="BYTE" END="413" FILEID="FID6" />
    </METS:fptr>
  </METS:div>
</METS:div>
</METS:div>
</METS:structMap>
<METS:structLink>
  <METS:smLink from="LINK1" to="PAGE2" xlink:title="Who Is She" />
  <METS:smLink from="LINK2" to="PAGE3" xlink:title="Agenda, Not Gender" />
  <METS:smLink from="LINK3" to="PAGE4" xlink:title="Favorites" />
  <METS:smLink from="LINK4" to="PAGE5" xlink:title="Photo Gallery" />
  <METS:smLink from="LINK5" to="PAGE6" xlink:title="Feedback" />
  <METS:smLink from="LINK6" to="PAGE1" xlink:title="Home" />
  <METS:smLink from="LINK7" to="PAGE3" xlink:title="Agenda, Not Gender" />
  <METS:smLink from="LINK8" to="PAGE4" xlink:title="Favorites" />
  <METS:smLink from="LINK9" to="PAGE5" xlink:title="Photo Gallery" />

```

```
<METS:smLink from="LINK10" to="PAGE6" xlink:title="Feedback"/>
<METS:smLink from="LINK11" to="PAGE1" xlink:title="Home"/>
<METS:smLink from="LINK12" to="PAGE2" xlink:title="Who Is She"/>
<METS:smLink from="LINK13" to="PAGE4" xlink:title="Favorites"/>
<METS:smLink from="LINK14" to="PAGE5" xlink:title="Photo Gallery"/>
<METS:smLink from="LINK15" to="PAGE6" xlink:title="Feedback"/>
<METS:smLink from="LINK16" to="PAGE1" xlink:title="Home"/>
<METS:smLink from="LINK17" to="PAGE2" xlink:title="Who Is She"/>
<METS:smLink from="LINK18" to="PAGE3" xlink:title="Agenda, Not Gender"/>
<METS:smLink from="LINK19" to="PAGE5" xlink:title="Photo Gallery"/>
<METS:smLink from="LINK20" to="PAGE6" xlink:title="Feedback"/>
<METS:smLink from="LINK21" to="PAGE1" xlink:title="Home"/>
<METS:smLink from="LINK22" to="PAGE2" xlink:title="Who Is She"/>
<METS:smLink from="LINK23" to="PAGE3" xlink:title="Agenda, Not Gender"/>
<METS:smLink from="LINK24" to="PAGE4" xlink:title="Favorites"/>
<METS:smLink from="LINK25" to="PAGE6" xlink:title="Feedback"/>
<METS:smLink from="LINK26" to="PAGE1" xlink:title="Home"/>
<METS:smLink from="LINK27" to="PAGE2" xlink:title="Who Is She"/>
<METS:smLink from="LINK28" to="PAGE3" xlink:title="Agenda, Not Gender"/>
<METS:smLink from="LINK29" to="PAGE4" xlink:title="Favorites"/>
<METS:smLink from="LINK30" to="PAGE5" xlink:title="Photo Gallery"/>
</METS:structLink>
</METS:mets>
```

APPENDIX 23

Sample Complex Data Structure Containing Metadata That Can Be Extracted from an arc File

What follows is a snippet from a series of hashes internal to an encompassing hash that structures the metadata for all of the websites in an .arc file. The .arc file in question is the April 17th crawl of the Nigerian Election sites. The key, on the left-hand side of the expression, contains the field information, and the value half of the expression contains the bit of metadata that we're interested in preserving. Material that was derived from the HTTP headers or the .arc file header for each file has its key in lower case. If metadata was derived from the <meta> tags in the HTML page archived in the .arc that field name is capitalized. This is the sort of data structure that could replace a database as a repository of information in field-value pairs. The tech team has chosen to walk this hash to create SQL INSERT statements to enter into a database repository. The SQL statements for the pages trotted out in hashes here is appended at the end of the document.

```
my %webpages = ();

$webpages{"0001http://www.nigeriancp.net:80/"} = ({"IP" => "209.157.71.50", "ArchiveDate" => "20030417223120",
"SortDate" => "20030417",
"Status" => "HTTP/1.1 200 OK",
"Server" => "Microsoft-IIS/5.0",
>Date" => " Thu, 17 Apr 2003 22:31:19 GMT",
"ContentType" => " text/html",
"LastModified" => " Tue, 18 Mar 2003 00:06:33 GMT",
"NormedLastModified" => "20030318",
"ContentLength" => " 37934",
"DESCRIPTION" => "after years of misrule from the likes of abacha, obasanjo, babangida, and the total collapse of our
economy, the ncp is here to rescue the masses..",
"KEYWORDS" =>
"nigeria, ncp, conscience, party, gani, fawehinmi, lanre, banjo, obasanjo, ibb, abacha, dele, giwa, falana, aborisade, waheed, alabe
de, ghana, africa, guardian, newspapers, newspaper, tribune, vanguard, amana, new, nigerian, alliance, democracy, human, rights
, watch, news, abia, adamawa, akwa, ibom, anambra, bauchi, bayelsa, benue, borno, cross, river, delta, ebonyi, ekiti, enugu, gombe, i
mo, jigawa, kaduna, kano, katsina, kebbi, birnin, kogi, kwara, lagos, nassarawa, niger, ogun, osun, oyo, plateau, poverty, eradication
, muslim, christian, conflict, tribal, politics, peoples, democratical, pdp, anpp, app, nec, national, electoral, world, commision, , govern
or, federal, government, ethnic, sokoto, taraba, yobe, zamfara, inec, bomb, soyinka, ken, saro, wiwa",
});
$webpages{"0002http://www.buhari2003.org:80/"} = ({"IP" => "63.251.4.139", "ArchiveDate" => "20030417223120",
"SortDate" => "20030417",
"Status" => "HTTP/1.1 200 OK",
>Date" => " Thu, 17 Apr 2003 22:31:20 GMT",
"Server" => " Apache/1.3.22 (Unix) (Red-Hat/Linux) ApacheJServ/1.1.2 mod_jk/1.2.0 mod_perl/1.2.4_01 PHP/4.2.2
FrontPage/5.0.2 mod_ssl/2.8.5 OpenSSL/0.9.6b",
"LastModified" => " Thu, 17 Apr 2003 20:32:47 GMT",
"NormedLastModified" => "20030417",
"ContentLength" => " 105569",
"ContentType" => " text/html",
});
$webpages{"0003http://www.homestead.com:80/~site/tool/Homestead/HC_Objects/Images/HCUser_Guestbook/simple/g
uestbook.gif"} = ({"IP" => "209.157.71.50", "ArchiveDate" => "20030417223120", "SortDate" => "20030417",
"Status" => "HTTP/1.1 200 OK",
"Server" => " Microsoft-IIS/5.0",
>Date" => " Thu, 17 Apr 2003 22:31:20 GMT",
"ContentType" => " image/gif",
"LastModified" => " Wed, 30 Oct 2002 23:43:43 GMT",
"NormedLastModified" => "20021030",
"ContentLength" => " 1818",
});
$webpages{"0004http://npgg.freecyberzone.com:80/"} = ({"IP" => "208.185.127.162", "ArchiveDate" =>
"20030417223120", "SortDate" => "20030417",
"Status" => "HTTP/1.1 200 OK",
>Date" => " Thu, 17 Apr 2003 22:31:20 GMT",
```

```

"Server" => ".V13 Apache/1.3.26 (Unix) mod_fs 6.005",
"LastModified" => " Wed Apr 9 00:27:09 2003 GMT",
"NormedLastModified" => "9Apr",
"ContentType" => " text/html",
"GENERATOR" => "Microsoft FrontPage 5.0",
});
$webpages{"0005http://www.muhammadubuhari.com:80/"} = ({"IP" => "64.91.233.115", "ArchiveDate" =>
"20030417223120", "SortDate" => "20030417",
"Status" => "HTTP/1.1 200 OK",
"Date" => " Thu, 17 Apr 2003 22:32:38 GMT",
"Server" => " Apache/1.3.27 (Unix) DAV/1.0.3 mod_throttle/3.1.2 mod_log_bytes/1.2 mod_bwlimited/1.0 PHP/4.3.1
FrontPage/5.0.2.2510 mod_ssl/2.8.12 OpenSSL/0.9.6b",
"ContentType" => " text/html",
"DESCRIPTION" => "Muhammadu Buhari",
"KEYWORDS" => "Muhammadu Buhari, Buhari, ANPP, 2003, elections, nigeria, presidential candidate",
});

```

A simple perl script walks the hash of a hash above and outputs the following series of SQL INSERT statements for loading the values into a MySQL database:

```

-- data for 0001http://www.nigeriancp.net:80/
INSERT INTO object (objID, url,
ContentType,KEYWORDS,NormedLastModified,Server,DESCRIPTION,IP,Status,ArchiveDate,LastModified,SortDate,Date,
ContentLength) VALUES (1, 'http://www.nigeriancp.net:80/', '
text/html','nigeria,ncp,conscience,party,gani,fawehinmi,lanre,banjo,obasanjo,ibb,abacha,dele,giwa,falana,aborisade,waheed
,alabede,ghana,africa,guardian,newspapers,newspaper,tribune,vanguard,amana,new,nigerian,alliance,democracy,human,ri
ghts,watch,news,abia,adamawa,akwa,ibom,anambra,bauchi,bayelsa,benue,bornu,cross,river,delta,ebonyi,ekiti,enugu,gomb
e,imo,jigawa,kaduna,kano,katsina,kebbi,birnin,kogi,kwara,lagos,nassarawa,niger,ogun,osun,oyo,plateau,poverity,eradication
,muslim,christian,conflict,tribal,politics,peoples,democratical,pdp,anpp,app,nec,national,electoral,world,commision,,governor
,federal,government,ethnic,sokoto,taraba,yobe,zamfara,inec,bomb,soyinka,ken,saro,wiwa','20030318',' Microsoft-
IIS/5.0','after years of misrule from the likes of abacha, obasanjo,babangida,and the total collapse of our economy, the ncp is
here to rescue the masses..','209.157.71.50','HTTP/1.1 200 OK','20030417223120',' Tue, 18 Mar 2003 00:06:33
GMT','20030417',' Thu, 17 Apr 2003 22:31:19 GMT',' 37934');
-- data for 0002http://www.buhari2003.org:80/
INSERT INTO object (objID, url,
ContentType,NormedLastModified,Server,IP,Status,ArchiveDate,LastModified,SortDate,Date,ContentLength) VALUES (2,
'http://www.buhari2003.org:80/', ' text/html','20030417',' Apache/1.3.22 (Unix) (Red-Hat/Linux) ApacheJServ/1.1.2
mod_jk/1.2.0 mod_perl/1.24_01 PHP/4.2.2 FrontPage/5.0.2 mod_ssl/2.8.5 OpenSSL/0.9.6b','63.251.4.139','HTTP/1.1 200
OK','20030417223120',' Thu, 17 Apr 2003 20:32:47 GMT','20030417',' Thu, 17 Apr 2003 22:31:20 GMT',' 105569');
-- data for
0003http://www.homestead.com:80/~site/tool/Homestead/HC_Objects/Images/HCTest/Guestbook/simple/guestbook.gif
INSERT INTO object (objID, url,
ContentType,NormedLastModified,Server,IP,Status,ArchiveDate,LastModified,SortDate,Date,ContentLength) VALUES (3,
'http://www.homestead.com:80/~site/tool/Homestead/HC_Objects/Images/HCTest/Guestbook/simple/guestbook.gif', '
image/gif','20021030',' Microsoft-IIS/5.0','209.157.71.50','HTTP/1.1 200 OK','20030417223120',' Wed, 30 Oct 2002 23:43:43
GMT','20030417',' Thu, 17 Apr 2003 22:31:20 GMT',' 1818');
-- data for 0004http://npgg.freecyberzone.com:80/
INSERT INTO object (objID, url,
ArchiveDate,LastModified,SortDate,ContentType,Date,NormedLastModified,Server,GENERATOR,IP,Status) VALUES (4,
'http://npgg.freecyberzone.com:80/', '20030417223120',' Wed Apr 9 00:27:09 2003 GMT','20030417',' text/html',' Thu, 17
Apr 2003 22:31:20 GMT','9Apr',' .V13 Apache/1.3.26 (Unix) mod_fs 6.005','Microsoft FrontPage
5.0','208.185.127.162','HTTP/1.1 200 OK');
-- data for 0005http://www.muhammadubuhari.com:80/
INSERT INTO object (objID, url, ArchiveDate,SortDate,ContentType,Date,KEYWORDS,Server,DESCRIPTION,IP,Status)
VALUES (5, 'http://www.muhammadubuhari.com:80/', '20030417223120','20030417',' text/html',' Thu, 17 Apr 2003 22:32:38
GMT','Muhammadu Buhari, Buhari, ANPP, 2003, elections, nigeria, presidential candidate',' Apache/1.3.27 (Unix) DAV/1.0.3
mod_throttle/3.1.2 mod_log_bytes/1.2 mod_bwlimited/1.0 PHP/4.3.1 FrontPage/5.0.2.2510 mod_ssl/2.8.12
OpenSSL/0.9.6b','Muhammadu Buhari','64.91.233.115','HTTP/1.1 200 OK');

```

APPENDIX 24

INSERT Statements Created from Extracting Metadata Out of the .dat File

A snippet of the .dat file that accompanies the .arc file examined above demonstrates that a module of the harvesting application has generated a .dat file header at the top of the file, a file header for each file collected, followed by a field-value pair for each piece of metadata that was filtered out of the .arc. The key to the field letters can be found in Appendix ***. Some items peculiar to this file are the checksums generated by the harvester and the listing of links distinguished by type: href to another page, src to an inline image, link to a script or to a .css page. In some cases website graphics are distinguished from inline images. This is also where the title is filtered out of the HTML page archived in the .arc.

```
<snip>
[ .dat file header ]
filedesc://disk/2003-04-17-15-31-crawl/NIGERIAN_ELLECTION_2003-030417153120-000.dat 0.0.0.0 20030417223120
text/plain 77
1 0 InternetArchive
URL IP-address Archive-date Content-type Archive-length

[ HTML file header ]
http://www.nigeriancp.net:80/ 209.157.71.50 20030417223120 alexa/dat 3210
[ field-value pairs ]
m text/html
s 200
c 935e815086f1326375add0fa15e8ea26
k 2b468e9b27db347017d709d30b4488f6
v 199
V 187
n 38163
t national conscience party of nigeria homepage,ncp
i www.nigeriancp.net/defaultUser/images/javascript_disabled.gif
i
uptpro.homestead.com/~site/Scripts_Track/track.dll?H_H=1750256636&H_P=100&H_A=0&H_V=2&H_I=1&H_U=600638
&E=51&E=616&E=613&E=33&E=457&E=8&E=461&E=39&E=491
i www.nigeriancp.net/files/flag2003.jpg
i www.nigeriancp.net/~site/Layout/TopImages/Black_Rectangle.gif
l www.nigeriancp.net/TCP.html
l www.nigeriancp.net/TCPemp.html
l www.nigeriancp.net/TCPfood.html
l www.nigeriancp.net/TCPwater.html
</snip>
```

A sequential reading of this file using perl can easily output SQL UPDATE statements to supplement the records that were entered previously from the .arc file dump.

```
-- data for 1http://www.nigeriancp.net:80/
UPDATE object set checksum1 = '935e815086f1326375add0fa15e8ea26' WHERE objID = 1;
UPDATE object set checksum2 = '2b468e9b27db347017d709d30b4488f6' WHERE objID = 1;
UPDATE object set title = 'national conscience party of nigeria homepage,ncp' WHERE objID = 1;
-- data for 2http://www.buhari2003.org:80/
UPDATE object set checksum1 = '61a43df1c350ff63279bfb6526473280' WHERE objID = 2;
UPDATE object set checksum2 = '1b3b7a75069ff40ad2260d4142370d8d' WHERE objID = 2;
UPDATE object set title = 'Buhari 2003' WHERE objID = 2;
-- data for
3http://www.homestead.com:80/~site/tool/Homestead/HC_Objects/Images/HCTest_Guestbook/simple/guestbook.gif
UPDATE object set checksum1 = 'a0d42547f0fa57e0fc3d7de22cf26bcc' WHERE objID = 3;
UPDATE object set checksum2 = 'a0d42547f0fa57e0fc3d7de22cf26bcc' WHERE objID = 3;
-- data for 4http://npgg.freecyberzone.com:80/
UPDATE object set checksum1 = '69cbdf29ef285c8f1306b4a5c1485c90' WHERE objID = 4;
UPDATE object set checksum2 = 'c9cbd78d178a792261e9bcb32515f2d6' WHERE objID = 4;
UPDATE object set title = 'NFGG' WHERE objID = 4;
```

```
-- data for 5http://www.muhammadubuhari.com:80/  
UPDATE object set checksum1 = '9eeb59b94848a824ecb30ac41d7b6070' WHERE objID = 5;  
UPDATE object set checksum2 = '351d6dba7730af1156528201e6510cb6' WHERE objID = 5;  
UPDATE object set title = 'Muhammadu Buhari' WHERE objID = 5;
```

The same script also outputs a load file to create a sorted linkTable containing one instance of each of the links contained in all the pages in the .arc, along with a linkLink table to create the many-to-many relationships between each page and its respective links.

The fields for the linkTable dump are linkName, the object that contains the link, and the linkType:

```
<snip>  
www.ndnigeria.com/index_files/filelist.xml|1|4  
www.ndnigeria.com/textstyle|1|4  
www.ndnigeria.com/textstyle.css|1|4  
www.ndnigeria.com/images/pc1.gif|1|2  
www.ndnigeria.com/nigfag.gif|1|2  
www.ndnigeria.com/images/pc2.gif|1|2  
www.ndnigeria.com/images/pc3.gif|1|2  
www.ndnigeria.com/b_pc42.jpg|1|2  
www.ndnigeria.com/images/pc5.gif|1|2  
</snip>
```


APPENDIX 25

Internet Archive arc Format and OAIS Metadata Framework

Nancy H. Holcomb, CTS Metadata Services

May 5, 2003

This document compares the Alexa Arc Format used by Internet Archive (IA) with the metadata framework of the OAIS Information Model. The relevant parts of the OAIS Information Model are those relating to Content Information and Preservation Description information.

Content Information comprises the content data object itself and its Representation Information.

Representation Information has two components: The **Content Data Object Description** comprising 13 metadata elements, and **Environment Description**. Environment Description comprises Software Environment and Hardware Environment. **Software Environment** includes two sections: Rendering Programs (two metadata elements with associated sub-elements) and Operating System (four metadata elements) **Hardware Environment** has three components: Computational Resources, Storage, and Peripherals. Under **Computational Resources** are three metadata elements; under **Storage** are two elements; under **Peripherals** are two elements. One element, Location, is listed under Hardware Environment as a Whole.

This report will list the metadata elements as given in the OAIS document, and discuss their location (or absence) in the Arc Format.

The Arc Format records have two parts: the version_block and the rest_of_arc_file specified as either <doc> or <doc><rest_of_arc_file>. The section labeled <doc> comprises <URL-record> and <network_doc>. The URL-record fields are identified in the last line of the version_block, called <URL-record-definition>.

The Version Block contains mainly information related to preservation description in IA.

The URL record gives name and size of an object in the Archive (the Content Data Object) and also gives some metadata about its retrieval.

The Archive-date in the URL record is the date the file was saved to the Archive, i.e., the Ingest date. The date in line 1 of Version Block is earlier than this. So it must be a pre-ingest date.

The rest of the Arc Format record, designated <network_doc> defined as “whatever the protocol returned,” is data gotten by the Web crawler that harvested the Object, including the Object itself. See questions at end re how this data fits in with OAIS.

Content Data Object Description

1. **Underlying abstract form description:** Human readable description of the underlying abstract form of the Content Data Object. In Arc Format, the 3rd line of the Version Block contains the URL Record Definition, a verbal description of the elements forming each URL Record in IA.
2. **Structural type:** Class of digital object to which the Content Data Object belongs. In <content-type> in URL record. content-type = "no-type" | MIME type of data (e.g., "text/html")
3. **Technical infrastructure of complex object:** Internal structure, i.e., list of components of a complex object and their interrelationships. Not present in Arc Format.
4. **File description:** Technical specifications of files comprising Content Data Object. Applies to file formats used to access content, rather than storage formats (e.g. ZIP files). Not present.
5. **Installation requirements:** Specialized procedures needed to install Content Data Object (hereafter designated "Object"). Not present.

6. **Size of Object (in bytes):** In the Arc Format, Archive-length is specified as the last element in the URL-record-definition. But in the list of elements for URL-record, <length> is given as the last element and defined as "ascii representation of size of network_doc in bytes." We're assuming that these two definitions are the same. In examples, within <network_doc>, there is a metadata element designated "Content-length" that seems to give the size in bytes of the HTML content.
7. **Access inhibitors:** Features of the Object intended to inhibit access. Not present.
8. **Access facilitators:** Methods to enhance access to information within Object that needs to be maintained. Not present.
9. **Significant properties:** Properties of the Object's content that must be preserved or maintained through successive cycles of preservation. Not present.
10. **Functionality:** Describes functional or "look and feel" attributes of Object in its current version in the archive. Not present.
11. **Description of rendered content:** Describes Object's content, as it should be viewed and interpreted by users. Not present.
12. **Quirks:** Loss of functionality or change in look and feel of Object resulting from preservation process and procedures done by archive. Not present.
13. **Documentation:** Supporting documentation necessary/useful to display and/or interpret Object. Includes sub-element: Location (URL) of documentation. Not present.

Environment Description: *None of the metadata elements discussed under Environment Description, either software or hardware, are overtly present in IA's Arc Format.*

Environment Description: Software Environment

1. **Rendering programs**
 - a. **Transformation process:** Description of process to transform byte stream of Object into Underlying Abstract Form.
 - i. **Transformer engine:** Name and version of specific software capable of carrying out the process described in the Transformation Process.

Sub-elements:

 - Parameters: Configurations needed on Transformer engine in order to have success
 - Input format: Format of Object worked on by Transformer engine--makes sure the two are compatible
 - Output format: Format produced as a result of processing Object with Transformer engine
 - Location: Description of where the Transformer engine can be obtained
 - Documentation: Documentation necessary or useful for operating the Transformer engine
 - Location of documentation: e.g. URL
 - b. **Display/Access application:** Software capable of displaying or accessing content of Object
 - i. Input format: Format of Object that Display/Access application works on
 - ii. Output format: Description of output expected from Display/Access application
 - iii. Location: Location of Display/Access application (where it can be obtained)

- iv. **Documentation:** Documentation necessary or useful for operating the Display/Access application
 - Location of documentation: e.g. URL

2. Operating system

- a. **OS name:** Software platform on which Rendering Programs operate
- b. **OS version:** Version of Operating System named above
- c. **Location:** Location of working copy of the OS specified in OS name and OS version
- d. **Documentation:** Documentation necessary or useful for operation of the OS specified
 - i. Location of documentation: e.g. URL

Environment Description: Hardware Environment

1. Computational resources

- a. **Microprocessor requirements:** Microprocessor specs necessary to operate Object's software environment
- b. **Memory requirements:** Memory necessary to operate Object's software environment
- c. **Documentation:** Supporting documentation needed for operation of the Computational resources
 - i. Location of documentation: e.g. URL

2. Storage

- a. **Storage information:** Permanent storage resources needed for operation of software environment and/or rendering Object
- b. **Documentation:** Supporting documentation needed for operation or use of Storage resources
 - i. Location of documentation: e.g. URL

3. Peripherals

- a. **Peripheral requirements:** Description of additional equipment needed to render or display the Object
- b. **Documentation:** Supporting documentation for operation or use of Peripherals
 - i. Location of documentation: e.g. URL

4. Hardware environment as a whole

- a. **Location:** Location of the physical devices needed to render the Object. Links Object to compatible Hardware Environment

Preservation Description Information: Four categories: Reference Information, Context Information, Provenance Information, and Fixity Information.

1. **Reference information:** Describes identification systems and identifiers both internal and external to the archive

- a. **Archival system identification:** Unique identification of the Object within the Archive. Sub-elements:

--**Value:** Alexa arc format files are named with a ".arc" extension, e.g. "IA-000001.arc." At the beginning of the version block the name includes prefix "filedesc://" for example, "filedesc://IA-001102.arc" but in URL record it appears only as <filename> "IA-001102.arc"

--**Construction method:** Describes how the Archival System Identification is created and assigned.

This explanation doesn't appear in the actual record for any file in arc format. The format itself says under Version Block: "The version block identifies the original filename, file version, and URL record fields of the archive file." "Original filename" is what is identified as the Value above.

--**Responsible agency:** Body that assigns and maintains the Archival System Identification. This is Internet Archive. "IA" appears in the file name under "value" above. Also, in Version Block, 2nd line, <origin-code> is defined as "Name of gathering organization with no white space." Examples specify this as "Alexa Internet."

- b. **Global identification:** Unique identification of Object external to the Archive. Sub-elements:

--**Value:** "The Archive format uses the standard URL specification to identify objects"--Arc format documentation. Example:

<http://www.dryswamp.edu:80/index.html>

This URL appears at the beginning of the URL Record in the Arc format. The last line of the Version Block is the <URL-record-definition> and as such it specifies URL as the first element of the URL Record.

--**Construction method:** How Global Identification is created and assigned. Not present (?)

--**Responsible agency:** Body that assigns and maintains the Global Identification. This also appears to be IA, although not explicit.

- c. **Resource description:** Information for resource discovery extracted from existing metadata, or created by archive to support its access functions. Sub-element:

--**Existing metadata:** Any metadata scheme used for the Object. May accompany Object on ingest or be discovered later. Not present in Arc format. Sub-element:

--**Existing records:** A single record of a metadata scheme describing the Content Date Object. Not present in Arc format.

2. **Context information:** Relationships of Content Information with its environment, including reasons for its creation and relations with other objects. The related objects can be either other manifestations of the Object itself, or other Objects whose intellectual content is related to the Object in question.

- a. **Reason for creation:** Information on why an Object was created. Not present in Arc format.

- b. **Relationships:** Gives significant relationships between this Object and other Objects. Not present. Sub-elements:

--**Manifestation:** Links to other versions of the Object, such as HTML and PDF, or versions in earlier versions of Microsoft Word. In IA, each record is assigned a version number, in 2nd line of the Version Block. Based on previous versions identified in the Archive. Sub-elements:

--**Relationship type:** Type of relationship between archived Object and another associated Object. Spells out-- "Manifestation in HTML; Manifestation in PDF" etc. Not overtly present in IA Arc format.

--**Identification:** Identifies the related Object, so as to link the two together. Not present in IA Arc format. Only version no. for each manifestation is present.

--**Intellectual content:** Documents relationships with other similar or related Objects. Sub-elements:

--**Relationship type**: Names the type of relationship between the two objects. Not present.

--**Identification of the related object**: Not present.

3. **Provenance information**: Gives history of Content Information: origin, changes to it over time, and chain of custody.

Each of the following metadata elements delineates a category of the chronology or life cycle of the Object. They are event-based, recording details of each event that fits in one of the categories.

- a. **Origin**: Describes process by which Object was created. Example: Scanning paper document at 600 dpi in TIFF format, storing it on CD-ROM
- b. **Pre-ingest**: Describes history of Object, in terms of maintenance, content changes, custody, etc., from origin to ingest into Archive.
- c. **Ingest**: Describes process of depositing (ingesting) Object into Archive. Example: Migrated to Archive's standard storage format; complex Object broken down into component parts; AIP(s) assembled.
- d. **Archival retention**: Describes maintenance, changes in content, management, etc., of Object during its retention in Archive.
- e. **Rights management**: Documents legal uses of Object while archived. Example: Access permissions, legal deposit responsibilities.

Each of the above categories is described by one or more event-based sets of metadata. Each **event** has the following set of sub-elements:

--**Designation**: Name of event being described.

--**Procedure**: Describes the procedure used to accomplish the event. Example: Describes timing and steps of a format migration.

--**Date**: Date event occurred establishes chronology of events.

--**Responsible agency**: Entity responsible for successful event occurrence.

--**Outcome**: Describes outcome of Event's latest occurrence. Example: Successfully migrated Object from MS 97 to PDF.

In Arc Format, the **first line of Version Block** has data describing adding a file to the Archive: <date> the Archive file was created, and <ip_address> of the machine that created the Archive file. These could go in the event "**Ingest**" above.

2nd line of Version Block contains <version-number>, <reserved>, and <origin-code>. Version number is Manifestation in Context information. Since IA doesn't define <reserved>, and the record examples show its value as "0", we can't place it in the OAIS context. But the origin code is given as Alexa Internet in examples.

The URL record includes Result-code, which documents the ingest process.

--**Designation**: Ingest.

--**Procedure**: Not present in Arc. Except perhaps Result-code above.

--**Date**: <date> from first line of Version Block

--**Responsible agency**: <origin-code> is Alexa Internet.

--**Outcome:** In OAIS this seems to be textual. In ARC Format, the ip_address from first line of Version Block defines the machine that created the Archive file. This implies a successful outcome; the ip_address could be placed here.

Additionally, Network_doc defined as "whatever the protocol returned," can include **pre-ingest** information, such as date last modified.

Archive records are created as each new version of the resource is ingested into the Internet Archive. Metadata is created to describe the file as it exists on ingest, including the version number. Is anything recorded about post-ingest checking on it, for content loss, etc.? I couldn't find anything there.

4. **Fixity information:** Data Integrity checks or Validation/Verification keys that ensure that the Object hasn't been altered without documentation of the change. This set of metadata ensures that the Object in the Archive matches its associated metadata. One element, with 4 sub-elements. Repeatable. New authentication metadata is required anytime an Event occurs that alters the bit stream of the Object or its associated metadata (?). If both pre-Event and post-event versions of the Object are retained by the Archive, then the old Authentication data will remain as part of the metadata associated with the earlier version.

- a. **Object authentication:** Authenticates the Object and its content. Makes sure that no undocumented change has occurred to the Object. Example: Digital signature, watermark, check sum. Sub-elements:

--**Authentication type:** Technique used to authenticate Object. Explicit description of authentication method. Example: Digital signature consisting of a 128-bit hash computed using MD5 one-way hash function, encrypted with a private key.

--**Authentication procedure:** Steps to implement Authentication Type, including pointers to documentation, software, etc. Example: Pointer to software capable of computing an MD5 hash (existing as another Object in the Archive)

--**Authentication date:** Date of most recent use of this Authentication Type. Establishes a temporal benchmark against which later manifestations or versions of the Object can be compared.

--**Authentication result:** Result of most recent archival use of this Authentication Type.

Fixity information in IA Arc Format:

--**Authentication type:** Several authentications are defined in the Format:

Length (in Version Block): Last element in 1st line of Version Block, length specifies, in bytes, the size of the rest of the version block.

Length (in URL record): ascii representation of size of network_doc in bytes. (<network_doc> is defined as "whatever the protocol returned")

Checksum (in URL record): ascii representation of a checksum of the data. The specifics of the checksum are implementation specific.

Offset (in URL record): offset in bytes from beginning of file to beginning of URL-record.

--**Authentication procedure:** Not explicitly mentioned in Arc Format.

--**Authentication date:** This is the most recent date the Archive used the specific authentication type.

--**Authentication result:** This could be several numbers, corresponding to the Authentication type(s) used.

See questions at end of PDI table on what's in the IA Arc Format that doesn't match anything in the OAIS metadata scheme.

APPENDIX 26

Internet Archive Arc Format and OAIS Metadata Framework Nancy Holcomb, CTS Metadata Services May 5, 2003

Comparison of Alexa Arc Format used by Internet Archive with metadata framework of the OAIS Information Model

Content Data Object: bit stream(s) being preserved. Appears in Arc Format as part of <network_doc> section, at least in some cases (examples of records).

Representation Information

Content Data Object Description element	In IA Arc?	Description of element; location (if present) in Arc Format
Underlying abstract form description	Yes	Human readable description of underlying abstract form of content data object. In Arc Format, URL-record definition (3 rd line of Version Block) spells out in words the elements forming the URL record for the Archive file
Structural type	Yes	Class of digital object. Present in <content-type> in URL record. Content-type = "no-type" MIME type of data (e.g., "text/html")
Technical infrastructure of complex object	No	Internal structure, i.e., list of components and their interrelationships
File description	No	Technical specs of file(s) comprising ... object
Installation requirements	No	Specialized procedures needed, if any
Size of object	Yes	Archive-length specified as last element in URL-record-definition. But, in list of elements for URL-record, <length> is given as last element and defined as "ascii representation of size of network_doc in bytes." We assume these two definitions are the same. In examples, within <network_doc>, there is a metadata element designated "Content-length" that seems to give size of HTML content in bytes.
Access inhibitors	No	Encryption, password protection, etc.
Access facilitators	No	Enhance access to information within object, e.g., navigational links in hypertext document
Significant properties	No	Must be preserved over time and cycles of preservation
Functionality	No	Functional or "look and feel" attributes of current instantiation

Description of rendered content	No	Describes object's content as viewed and interpreted by users
Quirks	No	Loss of functionality due to preservation processes or archival procedures, e.g., broken links
Documentation	No	Supporting documentation needed or useful to display and/or interpret object
Location of documentation	No	Such as URL

Environment Description: Software

Rendering Programs

Element name	In IA Arc?	Definition; location (if present) in IA Arc Format
<u>Transformation process</u>	No	Describes process to transform byte stream of Object into Underlying Abstract Form
Transformer engine	No	Name & version of specific software that will carry out Transformation process described above
Parameters	No	Runtime parameters configured on Transformer engine in order to have success
Input format	No	Format of Object worked on by Transformer engine--ensures compatibility between the two
Output format	No	Format produced by processing Object with Transformer engine
Location	No	Location of Transformer engine (URL)
Documentation	No	For operation of Transformer engine
Location		Of documentation, e.g. URL
Display/Access Application	No	Software capable of displaying or accessing content of Object
Input format	No	Description of format of Object that the Display/Access application works on
Output format	No	Description of output expected from Display/Access application
Location	No	Of Display/Access Application
Documentation	No	Necessary or useful to operate the Display/Access Application
Location		Of documentation, e.g., URL

Operating system

OS name	No	Software platform on which Rendering Programs operate
OS version	No	Version of Operating System above
Location	No	Of working copy of OS specified in OS name and OS version
Documentation	No	Necessary/useful for operation of specified OS
Location of documentation		e.g., URL

Environment Description: Hardware

Computational Resources

Element name	In IA Arc?	Definition; location (if present) in IA Arc Format
Microprocessor requirements	No	Specs needed to operate Object's software environment
Memory requirements	No	Memory necessary to operate Object's software environment
Documentation	No	Necessary/useful for operation/use of Computational Resources
Location of Documentation		e.g., URL
Storage information	No	Description of permanent storage resources necessary for operating software environment and/or rendering the Object, e.g., amount of hard disk space
Documentation	No	Supporting documentation needed for operation or use of Storage resources
Location of Documentation		e.g., URL
Peripheral requirements	No	Additional equipment needed to render/display Object
Documentation	No	Needed for operation or use of Peripherals
Location of Documentation		e.g., URL
Hardware environment as a whole: Location	No	Location of physical devices needed to render the Object

APPENDIX 27

Internet Archive Arc Format and OAIS Metadata Framework Preservation Description Information (PDI)

Nancy Holcomb
May 5, 2003

1. Reference Information

PDI Reference Information element	In IA Arc?	Definition; location (if present) in Arc Format
Archival System Identification (ASI)	Yes	Unique identification of Object within Archive
Value	Yes	Value given to identify the AIP. Each Alexa Arc format file has unique no. with ".arc" extension. ASI is at beginning of Version Block with a prefix, e.g., "filedesc://IA-001102.arc" and as <filename> in URL Record, e.g., IA-001102.arc.
Construction method	Not explicitly	Describes how ASI is created and assigned. Arc format says that the Version Block ...identifies original filename... which is "Value" above
Responsible agency	Yes	Body that assigns and maintains ASI. "IA" in file name. In Version Block, <origin-code> is "Name of gathering organization..." and this is "Alexa Internet" in examples
Global System Identification	Yes	Unique ID external to Archive. Repeatable.
Value	Yes	Standard URL, specified as first element in URL record
Construction method	No	How Global ID is created and assigned
Responsible agency	Yes	This is IA as above under ASI
Resource description	No	Resource discovery information either taken from existing metadata or created by Archive.
Existing metadata	No	Name of metadata scheme used for Object, may accompany it on ingest or be discovered later
Records	No	A single instantiation of a specific metadata scheme, associated with the Object

2. Context Information

PDI Context Information element	In IA Arc?	Definition; location (if present) in Arc Format
Reason for creation	No	Gives reason why Object was created
Relationships	Yes (?)	Documents relation of Object to other Objects. No specific documentation in Arc Format. But assigned Version number implies relationship
Manifestation	Yes (?)	Relation between this Object and its other manifestations. Version number is assigned to each new version of Object in IA.
Relationship type	Yes (?)	Type of relationship between Object and another manifestation of it. See Question 1 at end re version numbers
Identification	No (?)	ASI or GSI or link to bib record for related Object. IA only assigns integer to each new version of a resource ingested over time
Intellectual content	No	Documents other Objects with related content
Relationship type	No	Between two Objects with similar content
Identification	No	ASI or GSI or link to bib record for related Object

3. Provenance Information

PDI Provenance Information element	In IA Arc?	Definition; location (if present) in Arc Format
Origin	No	Describes process by which Object was created
Pre-Ingest	Yes (?)	Describes history of Object via series of Events, from creation to ingest into Archive. Arc format's <network_doc>, defined as "whatever the protocol returned" includes some pre-ingest information, such as date/time file was last modified. Also, date in Version Block could be pre-ingest date. See question re "Dates" at the very end.
Ingest: Event	Yes	Process by which Object is deposited into Archive
Procedure	No; except perhaps result-code in URL Record	Describes procedure used to accomplish event, in this case Ingest into Archive
Date of ingest	Yes	<date> from first line of URL Record
Responsible agency	Yes	<origin-code> in 2 nd line of Version Block is Alexa Internet
Outcome	Yes(?)	In 1 st line of Version Block, ip_address of machine that created Archive file implies successful outcome of Ingest Event.
Archival retention	No	Describes maintenance, changes in content, etc. of Object while retained in Archive
Rights management	No	Specifies legal uses of Object

4. Fixity Information

PDI Fixity Information element	In IA Arc?	Definition; location (if present) in Arc Format
Object authentication	Yes	Gives enough information to meet Archive's minimum requirements to authenticate Object and its content. In Arc Format, several types of Object Authentication are used. See below.
Authentication type	Yes	Method used to authenticate Object. Authentications defined in Arc format: Length (in Version Block): Last element in 1 st line of Version Block, length specifies, in bytes, the size of the rest of the version block. Length (in URL record): ascii representation of size of network_doc in bytes. (<network_doc> is "whatever the protocol returned") Checksum (in URL record): ascii representation of a checksum of the data. Implementation specific. Offset (in URL record): offset in bytes from beginning of file to beginning of URL-record
Authentication procedure	No	Procedural steps to implement Authentication Type, including pointers to supporting documentation, software, etc.
Authentication date	Yes	Most recent date Archive used Authentication type
Authentication result	Yes	Result of most recent use of Authentication type. In Arc format, could be several numbers corresponding to Authentication type(s) used.

Questions on elements in the IA Arc Format that are not present in the OAIS metadata scheme:

1. Version numbers in IA Arc imply relationship among instantiations of the same resource over time. Does OAIS address versions in the same format anywhere in its metadata framework?
2. "Reserved" area in Version Block: In the 2nd line of the Version Block, there's an element after the version number that is designated "reserved" and defined as "string with no white space." In the examples, it is given as "0" so we assume it is undefined and unused and not relevant to OAIS currently.
3. The Arc Format includes the IP address of the machine from which the source file was harvested, in URL Record, 2nd item. Where in OAIS is the source IP for the ingest file recorded? As a Content Data Object description element? Or in Provenance, as part of the Ingest procedure?
4. "Whatever the protocol returned" is the definition of <network_doc> in Arc Format. This section comes after the URL Record. It's what a crawl returns, including the file to be ingested. A list of its elements, taken from the example Arc record, follows. How does this information fit into OAIS?
 - a. HTTP/1.0
 - b. 200 (Result code)
 - c. Document follows
 - d. Date: Mon, 04 Nov 1996 14:21:06 GMT
 - e. Server: NCSA/1.4.1
 - f. Content-type: text/html Last-modified; Sat,10 Aug 1996 22:33:11 GMT
 - g. Content-length: 30
 - h. <HTML>Hello World!!!</HTML>
5. "Length" in Version Block: At the end of the first line of Version Block are two pieces of metadata that refer to the rest of the Version Block. "The content type of 'text/plain' simply refers to the remainder of the version block. The length specifies the size, in bytes, of the rest of the version block." Is this "length" a check or measure of the integrity of the metadata?
6. The Archive-date in the URL record is the date the file was saved to the Archive i.e. the Ingest date. The date in line 1 of Version Block is earlier than this. So it must be a pre-ingest date. How does this relate to OAIS?
7. Location. In URL Record <location> is defined as "-" or URL of re-direct. In the example, this is "-" in version 2, so does that mean that the location of the file as harvested the 2nd time was the same as it was the first time? The URL is the same for both versions. Wouldn't the URL show up the difference in location without this "-" or redirect? Is this location something different than the location of the actual file?

APPENDIX 28

A Survey of Robots.txt Exclusions for the CRL Testbed Sites Captured on 04 August 2003

Owners or managers of Web content can choose to protect Web-delivered material by writing a simple text file called a robots.txt exclusion and mounting it at Web server root. This protocol allows for the exclusion of selected robots, crawlers or agents from selected directories or pages, or in extreme cases it can exclude all agents from indexing any part of the site. Agent compliance is voluntary, and most well-intentioned agents will at least check for a robots.txt file before crawling a domain. In discussions between the curatorial and technical groups it was determined that whether we decide to comply with robots.txt rules or not, it would be useful to see what percentage of sites do protect material with a robots.txt exclusion, and what sorts of material is protected in particular. This survey is the result of a one-time crawl against the 635 URLs in the CRL testbed to determine how many sites use the robots.txt exclusion, and what sorts of exclusions exist.

Test runs brought to light a number of traps and problems that had to be accommodated in successive rewrites of the crawler. A number of the sites on our list had already disappeared (grist for the preservation mill); some servers timed out; some redirected the crawler to a custom 404 not found page for robots.txt and others delivered the default "not found" message and 404 page - the crawler had to be aware of both responses; some robots.txt pages were incorrectly written.

This simple perl LWP-based crawler was fed a list of domain-level URLs to crawl; it then tested for the existence of a robots.txt file by first filtering server response errors, and recording a 500 error (for server error) 404 (for no robots.txt page found). If there was an error-free response, the result was broken down into two categories: whether there was a custom 404 redirect which would mimic a normal HTML page for a crawler this simple, or whether an actual robots.txt file exists. The crawler wrote to a single report file, recording each server response and writing any extant robots.txt content into the file; it then tabulated the number of server errors, custom redirects and robots.txt pages for each region.

The following is a tabulation of how many sites by region had robots.txt files:

- Sub-Saharan Africa: 1 out of 64 1.5%
- South Asia: 2 out of 35 5.7%
- Latin America: 79 out of 476 16.5%
- Western Europe: 7 out of 60 11.5%

What the robots.txt exclusion protects on those sites that deploy it are for the most part cgi-bin scripting and image/graphics directories, along with directories containing resources such as javascripts and .css stylesheets. Also protected in a handful of cases were Web stats outputs, mail, logs and php admin directories.

Seven sites from this sample explicitly barred all crawlers from all directories. Around four others appear to have intended to bar all robots from all pages, but their syntax was incorrect. Interestingly, there were four cases of exclusions specifically against the ia_archiver, who provides us with our content:

<http://www.freemalaysia.com/robots.txt>
<http://www.clasemediia.org/robots.txt>
<http://www.ladeudaexterna.com/robots.txt>
<http://www.megaelecciones.com/robots.txt>

1. Sub-Saharan Africa

There were 59 pages with errors, of which 16 were 500 server errors; 4 redirects and 1 page with a robots.txt that could be captured.

<http://www.kenyaelections2002.org/robots.txt> has the following robot.txt:

```
User-agent: *  
  
Disallow: /_fpclass  
Disallow: /_private  
Disallow: /_themes  
Disallow: /_vti_cnf  
Disallow: /_vti_log  
Disallow: /_vti_pvt  
Disallow: /_vti_script  
Disallow: /_vti_txt  
Disallow: /cgi-bin  
Disallow: /email  
Disallow: /fpdb  
Disallow: /image  
  
#Disallow: /w3svc? #Change ? with the instance number please.
```

2. Southeast Asia

There were 30 pages with errors, of which 3 were 500 server errors; 3 redirects and 2 pages with a robots.txt that could be captured.

<http://www.abim.org.my/robots.txt> has the following robot.txt:

```
User-agent: *  
Disallow: admin.php  
Disallow: config  
Disallow: header  
Disallow: footer  
Disallow: pntables  
Disallow: referer  
Disallow: /images  
Disallow: /includes  
Disallow: /modules/NS-  
Disallow: /pnadodb  
Disallow: /themes
```

<http://www.freemalaysia.com/robots.txt> has the following robot.txt:

```
User-agent: *  
Disallow: /s  
Disallow: /c  
  
User-agent: ia_archiver  
Disallow: /  
  
User-agent: Scooter  
Disallow: /
```

3. Latin America

There were 356 pages with errors, of which 94 were 500 server errors; 41 redirects and 79 pages with a robots.txt that could be captured.

<http://ahorristas.8m.com/robots.txt> has the following robot.txt:

```
# Default /robots.txt File for FreeServers  
  
User-agent: *  
Disallow: /cgi-bin/
```

<http://ar.dir.yahoo.com/robots.txt> has the following robot.txt:

```
# Rover is a bad dog <http://www.roverbot.com>
User-agent: Roverbot
Disallow: /
```

<http://caceroleando.8m.com/robots.txt> has the following robot.txt:

```
# Default /robots.txt File for FreeServers

User-agent: *
Disallow: /cgi-bin/
```

<http://cartadocumento.8m.com/robots.txt> has the following robot.txt:

```
# Default /robots.txt File for FreeServers

User-agent: *
Disallow: /cgi-bin/
```

<http://ciberderecho.com.ar/robots.txt> has the following robot.txt:

```
# Default /robots.txt File for FreeServers

User-agent: *
Disallow: /cgi-bin/
```

<http://csd.queensu.ca/robots.txt> has the following robot.txt:

```
# go away
User-agent: *
Disallow: /
```

<http://economist.com/robots.txt> has the following robot.txt:

```
#
# Economist.com robots.txt
#
# Created MS 29 May 2001 Full disallow
# Amended MS 27 Jul 2001 Allow directories
# 21/07/2003 SR 90011 Allow Google but nothing else
#
User-agent: *
Disallow: /

User-agent: googlebot
Allow: /
Disallow: /search/
Disallow: /members/
Disallow: /subscriptions/
Disallow: /admin/
```

<http://espanol.clubs.yahoo.com/robots.txt> has the following robot.txt:

```
User-agent: *
Disallow:
```

<http://fp.chasque.apc.org:8081/robots.txt> has the following robot.txt:

```
User-agent: *
Disallow: /
```

<http://lanic.utexas.edu/robots.txt> has the following robot.txt:

```
# robots.txt for http://lanic.utexas.edu/

User-agent: *           # All spiders should avoid
Disallow: /cgi-bin/    # Script files
Disallow: /icons/     # Default icons
```

```
Disallow: /form/      # Form Conf
Disallow: /link/     # Link Verification
Disallow: /test/     # The test area for Web experimentation
```

<http://www.cen-prd.org.mx/robots.txt> has the following robot.txt:

```
user-agent: *
disallow: /
```

<http://pir.gq.nu/robots.txt> has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *
Disallow: /cgi-bin/
```

<http://rapacalp.com.ar/robots.txt> has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *
Disallow: /cgi-bin/
```

<http://razonyrevolucion.freeservers.com/robots.txt> has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *
Disallow: /cgi-bin/
```

<http://www.adelco.com.ar/robots.txt> has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *
Disallow: /cgi-bin/
```

<http://www.anf.org.br/robots.txt> has the following robot.txt:

```
User-agent: *
Disallow: /cgi-bin/
Disallow: /wusage
```

<http://www.apbyn.com.ar.nstempintl.com/robots.txt> has the following robot.txt:

```
# Default /robots.txt File for all Community Architect Partner pages
```

```
User-agent: *
Disallow: /cgi-bin/
```

<http://www.area.com.mx/robots.txt> has the following robot.txt:

```
User-agent: *
Disallow:
```

<http://www.argentinaxxi.8m.net/robots.txt> has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *
Disallow: /cgi-bin/
```

<http://www.asambleaalmagro.8m.com/robots.txt> has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *
Disallow: /cgi-bin/
```

<http://www.asambleabahiense.8k.com/robots.txt> has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```



```
User-agent: *  
Disallow: /cgi-bin/
```

http://www.asambleapaternal.4t.com/robots.txt has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *  
Disallow: /cgi-bin/
```

http://www.asambleawilde.8m.net/robots.txt has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *  
Disallow: /cgi-bin/
```

http://www.autodeterminacionylibertad.8k.com/robots.txt has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *  
Disallow: /cgi-bin/
```

http://www.bandera.org/robots.txt has the following robot.txt:

```
User-agent: WebClipping.com  
Disallow: /
```

```
User-agent: WebClipping  
Disallow: /
```

```
User-agent: WebClipping.com  
Disallow: /
```

```
User-agent: 209.73.228.163  
Disallow: /
```

```
User-agent: 209.73.228.167  
Disallow: /
```

```
User-agent: robot  
Disallow: /
```

```
User-agent: crawl  
Disallow: /
```

```
User-agent: spider  
Disallow: /
```

```
User-agent: *  
Disallow: /cgi-bin/  
Disallow: /admin/  
Disallow: /contenido/  
Disallow: /formularios/  
Disallow: /fotos/  
Disallow: /galeria/  
Disallow: /images/  
Disallow: /Library/  
Disallow: /lista/  
Disallow: /scripts/  
Disallow: /Templates/  
Disallow: /foros/  
Disallow: /_mmServerScripts/  
Disallow: /_notes/
```

```
Disallow: /Connections/  
Disallow: /webstats/
```

<http://www.barriosdepie.org.ar/robots.txt> has the following robot.txt:

```
User-agent: *  
Disallow: oneshare.html  
Disallow: fhad.php  
Disallow: /exit
```

<http://www.boliviahoy.com/robots.txt> has the following robot.txt:

```
# robots.txt  
User-agent: *  
Disallow: /include/  
Disallow: /themes/  
Disallow: /images/  
Disallow: /language/
```

<http://www.c-a-c-e-r-o-l-a-z-o.com.ar/robots.txt> has the following robot.txt:

```
# Allow robots to browse everywhere  
User-agent: *  
Disallow:
```

<http://www.causaresponde.s5.com/robots.txt> has the following robot.txt:

```
# Default /robots.txt File for FreeServers  
  
User-agent: *  
Disallow: /cgi-bin/
```

<http://www.cen-prd.org.mx/robots.txt> has the following robot.txt:

```
user-agent: *  
disallow: /
```

<http://www.ciudadpolitica.com/robots.txt> has the following robot.txt:

```
User-agent: *  
Disallow: /cgi-bin/  
Disallow: /tmp/  
Disallow: /cache/  
Disallow: /class/  
Disallow: /images/  
Disallow: /include/  
Disallow: /install/  
Disallow: /kernel/  
Disallow: /language/  
Disallow: /templates_c/  
Disallow: /themes/  
Disallow: /uploads/  
Disallow: /phpAdsNew/
```

<http://www.clasemedia.org/robots.txt> has the following robot.txt:

```
User-agent: *  
Disallow: /s  
Disallow: /c
```

```
User-agent: ia_archiver  
Disallow: /
```

```
User-agent: Scooter  
Disallow: /
```

<http://www.colatino.com/robots.txt> has the following robot.txt:

```
User-agent: *
```

```
Disallow:/cgi-bin
Disallow:/cgi-bin/link.cgi
Disallow:/cgi-bin/links.cgi
Disallow:/cgi-bin/Links.cgi
Disallow:/cgi-bin/Count.cgi
Disallow:/cgi-bin/ls1.pl
Disallow:/servlet/UserPrice
Disallow:/servlet/PrivateSale
Disallow:/servlet/buyDomDLS
Disallow:/jsp/search_pricerange.jsp
Disallow:/jsp/search_category.jsp
Disallow:/jsp/search_keyword.jsp
Disallow:/jsp/search_letter.jsp
Disallow:/ticker.jsp
```

http://www.comedor.8k.com/robots.txt has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *
Disallow: /cgi-bin/
```

http://www.contraelracismo.4t.com/robots.txt has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *
Disallow: /cgi-bin/
```

ERROR http://www.convocatoriaabierta.misionnet.com.ar/robots.txt: 401 Access Denied

http://www.copei.org/robots.txt has the following robot.txt:

```
# go away
User-agent: *
Disallow: /
```

http://www.deudoresnopesificados.8m.com/robots.txt has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *
Disallow: /cgi-bin/
```

http://www.elcacerolazo.org/robots.txt has the following robot.txt:

```
User-agent: *
Disallow: admin.php
Disallow: /admin/
Disallow: /images/
Disallow: /includes/
Disallow: /themes/
Disallow: /blocks/
Disallow: /manual/
Disallow: /modules/
Disallow: /language/
```

http://www.elistas.net/robots.txt has the following robot.txt:

```
User-agent: *
Disallow: /
```

http://www.enclaveroja.com.ar/robots.txt has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *
Disallow: /cgi-bin/
```

<http://www.escrache.com/robots.txt> has the following robot.txt:

```
User-agent: *  
Disallow: /webalizer/
```

<http://www.extraniolatino.20m.com/robots.txt> has the following robot.txt:
Default /robots.txt File for all Community Architect Partner pages

```
User-agent: *  
Disallow: /cgi-bin/
```

<http://www.farmonia.com.ar/robots.txt> has the following robot.txt:

```
User-agent: *  
Disallow: /cgi-bin/  
Disallow: /stats/
```

<http://www.fjcquilmes.8m.net/robots.txt> has the following robot.txt:
Default /robots.txt File for FreeServers

```
User-agent: *  
Disallow: /cgi-bin/
```

<http://www.frenteparaelcambio.org/robots.txt> has the following robot.txt:

```
User-agent: *  
Disallow: admin.php  
Disallow: /admin/  
Disallow: /images/  
Disallow: /includes/  
Disallow: /themes/  
Disallow: /blocks/  
Disallow: /modules/  
Disallow: /language/
```

<http://www.georgetown.edu/robots.txt> has the following robot.txt:

```
User-agent: *  
Disallow: /users*/maas475/  
Disallow: /anderson/  
Disallow: /home3/lewistc/www/  
Disallow: /Architext  
Disallow: /Architext.1.0.  
Disallow: /Architext.1.1.  
Disallow: /excite  
Disallow: /Excite  
Disallow: /lforte  
Disallow: /tamlit.old  
Disallow: /uis/systems/*  
Disallow: /acs  
Disallow: /uis/services/core/oldis/  
Disallow: /users/davisc/  
Disallow: /users/tuccyj/  
Disallow: /finaff/
```

<http://www.herramienta.com.ar/robots.txt> has the following robot.txt:

```
User-agent: *  
Disallow: admin.php  
Disallow: config  
Disallow: header  
Disallow: footer  
Disallow: pntables  
Disallow: referer  
Disallow: /images  
Disallow: /includes
```

```
Disallow: /modules/NS-  
Disallow: /pnadodb  
Disallow: /themes
```

<http://www.identidadsocialista.org.ar/robots.txt> has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *  
Disallow: /cgi-bin/
```

<http://www.juninytucuman.8m.com/robots.txt> has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *  
Disallow: /cgi-bin/
```

<http://www.juventudguevarista.8m.com/robots.txt> has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *  
Disallow: /cgi-bin/
```

<http://www.lacacerola.com/robots.txt> has the following robot.txt:

```
User-agent: appie  
Disallow: /
```

```
User-agent: *  
Disallow:
```

<http://www.ladeudaexterna.com/robots.txt> has the following robot.txt:

```
User-agent: *  
Disallow: /s  
Disallow: /c
```

```
User-agent: ia_archiver  
Disallow: /
```

```
User-agent: Scooter  
Disallow: /
```

<http://www.lanacion.com.ar/robots.txt> has the following robot.txt:

```
# Robots.txt (archivo)  
User-agent: *  
Disallow: /acumulados  
Disallow: /administracion  
Disallow: /anexos  
Disallow: /css  
Disallow: /edicionesanteriores #?????  
Disallow: /espec #?????  
Disallow: /herramientas  
Disallow: /imgs  
Disallow: /pruebas  
Disallow: /scripts  
Disallow: /servicio  
Disallow: /styles  
Disallow: /ustedOpina  
Disallow: /varios  
Disallow: /wap  
Disallow: /02 #ediciones anteriores  
Disallow: /01  
Disallow: /00  
Disallow: /99
```

Disallow: /98
Disallow: /97
Disallow: /96
Disallow: /95
Disallow: /suples/scripts
Disallow: /suples/styles
Disallow: /suples/varios
Disallow: /suples/arquitectura/02
Disallow: /suples/arquitectura/01
Disallow: /suples/arquitectura/00
Disallow: /suples/arquitectura/99
Disallow: /suples/arquitectura/98
Disallow: /suples/arquitectura/97
Disallow: /suples/arquitectura/96
Disallow: /suples/arquitectura/95
Disallow: /suples/arte/02
Disallow: /suples/arte/01
Disallow: /suples/arte/00
Disallow: /suples/arte/99
Disallow: /suples/arte/98
Disallow: /suples/arte/97
Disallow: /suples/arte/96
Disallow: /suples/arte/95
Disallow: /suples/autos/02
Disallow: /suples/autos/01
Disallow: /suples/autos/00
Disallow: /suples/autos/99
Disallow: /suples/autos/98
Disallow: /suples/autos/97
Disallow: /suples/autos/96
Disallow: /suples/autos/95
Disallow: /suples/campo/02
Disallow: /suples/campo/01
Disallow: /suples/campo/00
Disallow: /suples/campo/99
Disallow: /suples/campo/98
Disallow: /suples/campo/97
Disallow: /suples/campo/96
Disallow: /suples/campo/95
Disallow: /suples/ccioext/02
Disallow: /suples/ccioext/01
Disallow: /suples/ccioext/00
Disallow: /suples/ccioext/99
Disallow: /suples/ccioext/98
Disallow: /suples/ccioext/97
Disallow: /suples/ccioext/96
Disallow: /suples/ccioext/95
Disallow: /suples/cocina/02
Disallow: /suples/cocina/01
Disallow: /suples/cocina/00
Disallow: /suples/cocina/99
Disallow: /suples/cocina/98
Disallow: /suples/cocina/97
Disallow: /suples/cocina/96
Disallow: /suples/cocina/95
Disallow: /suples/cultura/02
Disallow: /suples/cultura/01
Disallow: /suples/cultura/00
Disallow: /suples/cultura/99
Disallow: /suples/cultura/98
Disallow: /suples/cultura/97
Disallow: /suples/cultura/96
Disallow: /suples/cultura/95

Disallow: /suples/empleos/02
Disallow: /suples/empleos/01
Disallow: /suples/empleos/00
Disallow: /suples/empleos/99
Disallow: /suples/empleos/98
Disallow: /suples/empleos/97
Disallow: /suples/empleos/96
Disallow: /suples/empleos/95
Disallow: /suples/encasa/02
Disallow: /suples/encasa/01
Disallow: /suples/encasa/00
Disallow: /suples/encasa/99
Disallow: /suples/encasa/98
Disallow: /suples/encasa/97
Disallow: /suples/encasa/96
Disallow: /suples/encasa/95
Disallow: /suples/enfoques/02
Disallow: /suples/enfoques/01
Disallow: /suples/enfoques/00
Disallow: /suples/enfoques/99
Disallow: /suples/enfoques/98
Disallow: /suples/enfoques/97
Disallow: /suples/enfoques/96
Disallow: /suples/enfoques/95
Disallow: /suples/infor/02
Disallow: /suples/infor/01
Disallow: /suples/infor/00
Disallow: /suples/infor/99
Disallow: /suples/infor/98
Disallow: /suples/infor/97
Disallow: /suples/infor/96
Disallow: /suples/infor/95
Disallow: /suples/revista/02
Disallow: /suples/revista/01
Disallow: /suples/revista/00
Disallow: /suples/revista/99
Disallow: /suples/revista/98
Disallow: /suples/revista/97
Disallow: /suples/revista/96
Disallow: /suples/revista/95
Disallow: /suples/teve/02
Disallow: /suples/teve/01
Disallow: /suples/teve/00
Disallow: /suples/teve/99
Disallow: /suples/teve/98
Disallow: /suples/teve/97
Disallow: /suples/teve/96
Disallow: /suples/teve/95
Disallow: /suples/turismo/02
Disallow: /suples/turismo/01
Disallow: /suples/turismo/00
Disallow: /suples/turismo/99
Disallow: /suples/turismo/98
Disallow: /suples/turismo/97
Disallow: /suples/turismo/96
Disallow: /suples/turismo/95
Disallow: /suples/vialibre/02
Disallow: /suples/vialibre/01
Disallow: /suples/vialibre/00
Disallow: /suples/vialibre/99
Disallow: /suples/vialibre/98
Disallow: /suples/vialibre/97
Disallow: /suples/vialibre/96

Disallow: /suples/vialibre/95

http://www.laprensahn.com/robots.txt has the following robot.txt:

```
User-agent: *
Disallow: /anuncios/
Disallow: /banners/
Disallow: /beta/
Disallow: /analog/
Disallow: /como_files/
Disallow: /miad/
Disallow: /panel/
Disallow: /test/
Disallow: /testing/
```

http://www.lavozdelconsumidor.com.ar/robots.txt has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *
Disallow: /cgi-bin/
```

http://www.lopezmurphy.com/robots.txt has the following robot.txt:

```
User-agent: *
Disallow: /CSS/
Disallow: /difundir/
Disallow: /HOME/
Disallow: /IMAGES/
Disallow: /interactuar/
Disallow: /JS/

Disallow: /TOOLS/
Disallow: /lopezmurphy_archivos/
Disallow: /lopezmurphy_files/
```

http://www.martinpresidente.com/robots.txt has the following robot.txt:

```
User-agent: *
Disallow: oneshare.html
Disallow: fhad.php
Disallow: /exit
```

http://www.megaelecciones.com/robots.txt has the following robot.txt:

```
User-agent: *
Disallow: /s
Disallow: /c

User-agent: ia_archiver
Disallow: /

User-agent: Scooter
Disallow: /
```

http://www.mycgiserver.com/robots.txt has the following robot.txt:

```
User-agent: *
Disallow: /images/
Disallow: /js/
Disallow: /exec/
Allow: /pdf/mcs-upgrade-infra.pdf
```

http://www.nacion.com/robots.txt has the following robot.txt:

```
User-agent: *
Disallow: /ancora/1996
Disallow: /ancora/1997
Disallow: /ancora/1998
```



```
Disallow: /ancora/1999
Disallow: /ancora/2000
Disallow: /ancora/2001
Disallow: /dominical/1996
Disallow: /dominical/1997
Disallow: /dominical/1998
Disallow: /dominical/1999
Disallow: /dominical/2000
Disallow: /dominical/2001
Disallow: /enforma/1996
Disallow: /enforma/1997
Disallow: /enforma/1998
Disallow: /enforma/1999
Disallow: /enforma/2000
Disallow: /enforma/2001
Disallow: /ln_ee/1995
Disallow: /ln_ee/1996
Disallow: /ln_ee/1997
Disallow: /ln_ee/1998
Disallow: /ln_ee/1999
Disallow: /ln_ee/2000
Disallow: /ln_ee/2001
Disallow: /moda/2000
Disallow: /moda/2001
Disallow: /tiempolibre/1998
Disallow: /tiempolibre/1999
Disallow: /tiempolibre/2000
Disallow: /tiempolibre/2001
Disallow: /viva/1997
Disallow: /viva/1998
Disallow: /viva/1999
Disallow: /viva/2000
Disallow: /viva/2001
Disallow: /zurqui/1996
Disallow: /zurqui/1997
Disallow: /zurqui/1998
Disallow: /zurqui/1999
Disallow: /zurqui/2000
Disallow: /zurqui/2001

Disallow: /archivo_ca
Disallow: /cgi-bin/
```

<http://www.observatorioelectoral.org/robots.txt> has the following robot.txt:

```
User-agent: *
Disallow: /PHP/
Disallow: /PHORUM/
Disallow: /AQ/
Disallow: /IMG/
Disallow: /ERRORS/
Disallow: /CSS/
Disallow: /FILES/
Disallow: /FRAMES/
Disallow: /UTILITIES/
Disallow: /JS/
Disallow: /LOGs/
```

<http://www.patrialibre.org.ar/robots.txt> has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *
Disallow: /cgi-bin/
```

<http://www.prt.5u.com/robots.txt> has the following robot.txt:

```
# Default /robots.txt File for all Community Architect Partner pages
```

```
User-agent: *  
Disallow: /cgi-bin/
```

<http://www.psecuador.org/robots.txt> has the following robot.txt:

```
User-agent: * # directed to all spiders, not just Scooter
```

```
Disallow: /cgi-bin/  
Disallow: /imgs/
```

<http://www.psecuador.org//robots.txt> has the following robot.txt:

```
User-agent: * # directed to all spiders, not just Scooter
```

```
Disallow: /cgi-bin/  
Disallow: /imgs/
```

<http://www.ptbrs.org.br/robots.txt> has the following robot.txt:

```
User-agent: *  
Disallow: /cgi-bin/  
Disallow: /wusage
```

<http://www.reclamosweb.com.ar/robots.txt> has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *  
Disallow: /cgi-bin/
```

<http://www.rincondominicano.com/robots.txt> has the following robot.txt:

```
User-agent: *  
Disallow: /cgi-bin  
Disallow: /phpBB2  
Disallow: /provincias/phpBB2  
Disallow: /gallery  
Disallow: /clasificados  
Disallow: /counter  
Disallow: /felix  
Disallow: /ksearch  
Disallow: /lapizarra  
Disallow: /lib  
Disallow: /luisvargas  
Disallow: /netpbm  
Disallow: /sendcard  
Disallow: /serversecure  
Disallow: /tv  
Disallow: /wusage  
Disallow: /resume  
Disallow: /icon  
Disallow: /images  
Disallow: /language_files  
Disallow: /iB_html  
Disallow: /graficas  
Disallow: /albums  
Disallow: /_private  
Disallow: /z
```

```
#idiot-bot; generates tons of 404 and malformed urls
```

```
User-agent: wget  
Disallow: /
```

```
User-agent: ExtractorPro
```

Disallow: /

http://www.simeca.org.ar/robots.txt has the following robot.txt:

Default /robots.txt File for FreeServers

User-agent: *

Disallow: /cgi-bin/

http://www.solucioncorralito.8m.com/robots.txt has the following robot.txt:

Default /robots.txt File for FreeServers

User-agent: *

Disallow: /cgi-bin/

http://www.superzebra.om3.net/robots.txt has the following robot.txt:

go away

User-agent: *

Disallow: /

http://www.terranova.50megs.com/robots.txt has the following robot.txt:

Default /robots.txt File for all Community Architect Partner pages

User-agent: *

Disallow: /cgi-bin/

http://www.transparency.org/robots.txt has the following robot.txt:

robots.txt file

#

Tells robots to where to get off.-)

#

For more info consult:

<http://info.webcrawler.com/mak/projects/robots/norobots.html>

#

Hacker wannabes, take note: there are no pointers here to any sensitive

stuff. Don't believe me? Well, if you want

to waste your time, don't let me stop you.

Treat all Robots equally...

User-agent: *

Disallow: /cgi-bin/

Robots have no business in here.

Disallow: /dnld/

Disallow: /gb_archive/

Disallow: /images/

Disallow: /photogallery/

Disallow: /poll/

Disallow: /order_shirt.html

Disallow: /t-shirt.html

Disallow: /notfound.php

Disallow: /guestbook.html

Disallow: /oneworld

Disallow: /ti-cir

http://www.uca.edu.sv/robots.txt has the following robot.txt:

robots.txt for http://www.uca.edu.sv

User-agent: *

Disallow: /home/fernandz/public_html/

http://www.uol.com.br/robots.txt has the following robot.txt:

User-Agent:

Disallow:

http://www.villaalba.8m.com/robots.txt has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *  
Disallow: /cgi-bin/
```

http://www.web1x1.org/robots.txt has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *  
Disallow: /cgi-bin/
```

http://www.webklan.com/robots.txt has the following robot.txt:

```
User-agent: *  
Disallow: admin.php  
Disallow: /admin/  
Disallow: /images/  
Disallow: /includes/  
Disallow: /themes/  
Disallow: /blocks/  
Disallow: /modules/  
Disallow: /language/
```

http://www.webpopular.4t.com/robots.txt has the following robot.txt:

```
# Default /robots.txt File for FreeServers
```

```
User-agent: *  
Disallow: /cgi-bin/
```

4. Western Europe

There were 48 pages with errors, of which 1 was a 500 server error; 5 redirects and 7 pages with a robots.txt that could be captured.

http://nopasaran.samizdat.net/robots.txt has the following robot.txt:

```
User-agent: *  
Disallow: *
```

http://www.alternativelibertaire.org/robots.txt has the following robot.txt:

```
User-agent: *
```

```
Disallow:
```

http://www.cgt.es/robots.txt has the following robot.txt:

```
User-agent: *  
Disallow: admin.php  
Disallow: /admin/  
Disallow: /images/  
Disallow: /includes/  
Disallow: /themes/  
Disallow: /blocks/  
Disallow: /modules/  
Disallow: /language/
```

http://www.gisti.org/robots.txt has the following robot.txt:

```
# Fichier lu par les moteurs de recherche.  
# Interdit l'accès au dossier 'gistinet'
```

```
User-agent: *  
Disallow: /gistinet
```

<http://www.le-pressoir.com/robots.txt> has the following robot.txt:

```
User-agent: *  
Disallow: oneshare.html  
Disallow: fhad.php  
Disallow: /exit
```

<http://www.lutte-ouvriere.org/robots.txt> has the following robot.txt:

```
# Robots.txt file  
  
User-agent: *  
Disallow: /cgi-bin  
Disallow: /stats  
Disallow: /stats_old  
Disallow: /gra  
Disallow: /z-tk
```

<http://www.pcf.fr/w2/robots.txt> has the following robot.txt:

```
User-agent: *  
Disallow: /mail/  
Disallow: /Library/  
Disallow: /Templates/  
Disallow: /images/  
Disallow: /prive/  
Disallow: /demo/
```

APPENDIX 29

<META> Tag Usage for Western European & Nigerian Election Sites

This survey was undertaken to determine which sites' homepages contain <meta> tags and <title> tags that could be harvested for descriptive or technical metadata. HTML <meta> tags appear in the head of an HTML document, and are thus hidden from the user's view. They were originally intended to aid indexing agents in categorizing, searching and ranking sites. The content of <meta> tags is machine or creator-generated descriptive and technical metadata ranging from a description of the page and relevant keywords to character set and encoding information, and the software and platform that created the page. In the early days search engines such as AltaVista could be tricked into ranking any given page higher through certain manipulations of the <meta> tag values; this practice of <meta> tag abuse called *spamdexing* continues in spite of the fact that current crawlers use much more sophisticated means for indexing and ranking.

This evaluation was undertaken before we had found a means of directly evaluating material in the .arc files. For this study a simple perl LWP agent was fed the list of all of the testbed URLs that had been crawled by the Alexa/IA crawler; the survey presented here is a representative sampling that takes into consideration 26 Western European and 20 Nigerian Election testbed URLs. The perl agent crawled the homepage of each site to collect <meta> values and write them to a report; it also tabulated how many sites did not use descriptive <meta> tags, and how many did not use any tags at all.

The result set reveals that a little more than half of the European sites (very small sampling, admittedly) used meta tags. On the Nigerian Election side 75% of the sites had no descriptive metadata.

A perusal of the content of the tags will confirm that the value of the descriptive metadata generated by the webpage creator where it occurs is certainly questionable, but is in itself something to be archived. Even the title page, where the desire to trick crawlers into ranking a page more highly does not come into play, can be extremely untrustworthy. In this sample titles were largely missing. On the other hand, machine-generated and human-generated technical metadata may be more trustworthy. Items that can be collected from this sampling include the generating software and operating system for the page (e.g. Adobe Page Mill 3.0 mac) and the character set for the page.

An interesting study for a future date would be to adapt the crawler to recognize where meta tags are incorrectly formatted, since page scrapers or meta parsers might be adversely affected by badly tagged <meta>s. A nagging problem that we have encountered in using this material is the use of Microsoft codepage and other non-Unicode characters; one site, <http://www.ac.eu.org/>, even mixes html entities and MS codepage characters. Another site, <http://www.gisti.org/>, provided bilingual versions of the description and keywords tags.

What follows is a listing of the homepage crawled followed by the <meta> tags and <title> tag harvested. Of the 26 European sites, 9 had no tags at all; 7 had no descriptive; and 10 had no title. For the Nigerian Election sites 14 had no tags, 3 no descriptive and 4 no title.

European Sites

1. http://www.ldmcom.org/adel_spartacus/
 <META NAME="DESCRIPTION" CONTENT="L'association éditrice A.D.E.L se consacre à l'édition et à la diffusion des textes des différentes tendances de la fraction gauche historique du mouvement communiste international.">
 <META NAME="KEYWORDS" CONTENT="ultra - gauche, révolution, prolétariat, luttes de classes, Dauvé, Barrot, Mattick, Korsch, Pannekoek, matérialisme, dialectique, Espagne 36, situationnisme, Debord, mouvement communiste, communauté humaine, Bordiga, gauche allemande, gauche italienne, internationalisme, Bricianer, Marx, Rubel, marxisme, militantisme, anticapitalisme, anarchisme, libertaire, communisme, bolchevisme, ">
 <TITLE>Association Documentation Edition Liaisons</TITLE>
2. <http://web.tiscali.it/anticitoyennisme/>
 NONE

3. <http://cettesemaine.free.fr/>
NONE
4. <http://www.federation-anarchiste.org/ml/>

```
<meta name="description"
content="Menu principal du site de la Fédération anarchiste">
<meta name="keywords"
content="fédération,anarchiste,FA,anarchisme,anarchie,communisme,libertaire">
<meta name="author" content="nono rude doodle hooligan">
```
5. <http://nopasaran.samizdat.net/>

```
<META NAME="GENERATOR" CONTENT="Adobe PageMill 3.0 Win">
= NO DESCRIPTIVE
```
6. <http://abirato.free.fr/3oiseau/OISEAU.HTM>

```
<meta http-equiv="content-type" content="text/html;charset=iso-8859-1">
<meta name="generator" content="Adobe GoLive 4">
= NO DESCRIPTIVE
```
7. <http://www.le-presseoir.com/>

```
<META http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<META name="COPYRIGHT" content="Anticopyright">
<META name="GENERATOR" content="StarOffice/5.2 (Win32)">
<META name="AUTHOR" content="WebMaster, Didier SOMVONGS">
<META name="CREATED" content="20000920">
<META name="CHANGED" content="20020611">
<META name="CLASSIFICATION" content="moissac, tarn-et-garonne, france">
<META name="DESCRIPTION" content="Journal d'idées libertaire égalitaire fraternelle">
<META name="KEYWORDS" content="libertaire, égalitaire, fraternitaire, anarchiste, anarchisme, situationniste,
Situationist International, situationists, situationism,internationale situationniste, situationnistes, situationnisme,
anarchism, anarchist, anarchy, littérature, philosophie, politique, poésie, poèmes, poesie, poemes, prose, journal,
journaux, quotidien, quotidiens, hebdomadaire, revue, presse, dada, dadaïste, dadaïsme, dadaïst, dadaïst">
<META name="Le Presseoir" content="Page d'accueil">
<META name="ROBOTS" content="index, follow">
<META name="OWNER" content="le-presseoir.com">
<META name="REPLY-TO" content="redaction@le-presseoir.com">
```
8. <http://reflexes.samizdat.net/>

```
<meta name="author" content="samizdat.net Pedro">
<meta name="ROBOTS" content="all, follow, index">
<meta name="DESCRIPTION" content="REFLEXes magazine antifasciste radical">
```
9. <http://tranquillou.free.fr/>
NONE
10. <http://membres.lycos.fr/tempcritiques/>

```
<META HTTP-EQUIV="Content-Type" CONTENT="text/html;charset=ISO-8859-1">
<META NAME="AUTHOR" CONTENT="TC">
<META NAME="KEYWORDS" CONTENT="">
<META NAME="DESCRIPTION" CONTENT="Site de la revue Temps Critiques">
```
11. <http://www.geocities.com/Paris/Opera/3542/>
NONE
12. <http://troploin0.free.fr/>

```
<META content="text/html; charset=windows-1252" http-equiv=Content-Type>
<META content="MSHTML 5.00.2314.1000" name=GENERATOR>
<META content="La révolution ne sera ni prise du pouvoir ni autogestion du salariat, mais émergence d'une communauté humaine qui n'aura pas besoin d'argent, de travail, d'identités, d'Etat." name=description>
<META content=" troploin, révolution, communisme, mouvement communiste, prolétariat, classe ouvrière, abolition du salariat, travail, communauté humaine, morale, révolution russe, révolution allemande, espagne 36, gauche allemande, gauche italienne, conseillisme, démocratie, fascisme, Marx, Bordiga, Pannekoek, Camatte, internationale situationniste,
```

Jean Barrot, Gilles Dauvé, La Banquise, Vieille Taupe, Mattick, Korsch, lutte de classe, communisme de conseil, marxisme, " name=keywords>

13. <http://www.cnt-f.org/>

```
<meta http-equiv="content-type" content="text/html; charset=iso-8859-1">
<title>Confédération Nationale du Travail (CNT)</title>
<meta name="keywords" content="CNT, cnt, politique, anarchy, anarchist, anarchie, anarchisme, anarchismo,
anarcho, syndicat, syndicalisme, trade-union, union, trade, trade-unionism anarcho-syndicalisme, AIT, IWA,
underground, lutte, lucha, struggle, policy, politics, polity, political, politica, politische, France, french, Francia,
Frankreich, Espagne, guerre civile, travail, work, labor, workers, travailleurs, Arbeit, trabajo, trabadores, internacional,
internacional, transnational, Proudhon, Bakounine, English, Spanish, espanol, Europe, antifa, antifascisme, feminism,
féminisme, anarcha, action directe, autogestion, direct action, self management, démocratie, democracy, direct,
directo, directa, éducation, education, université, universidad, FAU, social, sozial, accion, extrême, extreme, left,
gauche, Links, révolutionnaire, revolution, revolutionary, revolucion, revolucionario, solidarité, solidarity, solidaridad,
utopie, utopy, utopia, alternatif, alternativte, combat, fight, liberté, liberty, libertad, libertaire, libertarianism, espana,
Ferrer">
<meta name="description" content="Site de la Confédération Nationale du Travail (CNT),
syndicat révolutionnaire et anarcho-syndicaliste - Site of the CNT,
revolutionary and anarcho-syndicalist trade-union :
international struggles, direct action, self management, direct democracy, ...">
```

14. <http://www.confederationpaysanne.fr/>

NONE

15. <http://www.fsu.fr/>

```
<meta http-equiv="Content-Type" content="text/html; charset=windows-1252">
<meta name="GENERATOR" content="Microsoft FrontPage 4.0">
<meta name="ProgId" content="FrontPage.Editor.Document">
<title>Bienvenue sur le site de la FSU</title>
<meta name="Microsoft Theme" content="fusion-fond-yl 111">
<meta name="Microsoft Border" content="none">
= NO DESCRIPTIVE
```

16. <http://www.force-ouvriere.fr/>

```
<meta http-equiv="content-type" content="text/html; charset=iso-8859-1">
<meta name="generator" content="Adobe GoLive 5">
<title>FORCE OUVRIERE</title>
```

17. <http://www.g10.ras.eu.org/>

```
<meta name="Author" content="Solidaires union syndicale G10">
<meta name="Description" content="page accueil de Solidaires union syndicale G10">
<meta http-equiv="content-language" content="fr">
<meta name="Keywords" content="solidaires, SOLIDAIRES, solidaire, SOLIDAIRE, syndicat, Syndicat, SYNDICAT, G10,
g10">
```

18. <http://www.ac.eu.org/>

```
<meta name="Description" content="Page d'accueil du site des collectifs d'AC!, association de lutte contre le chômage, la
précarité et toutes les formes d'exclusion">
<meta name="Keywords" content="Agir ensemble contre le chômage, AC!, site des collectifs AC!,pr&eacute;carit&eacute;e,
exclusions, exclusion, droit au transport, edf, taxe d'habitation, logement, anti-expulsion, r&eacute;duction du temps de
travail, CCAS, ANPE, ASSEDIC, UNEDIC, allocations chômage, RMI, revenu minimum d'insertion, RMA, revenu minimum
d'activité, insertion, PARE, PAP, accompagnement, securitaire, gratuité, services publics, sans papiers, controle social,
antiseuritaire, g8, evian, porto alegre, seville, bruxelles, criminalisation, pauvrete, misere, exploitation, plein emploi,
st69">
<title>BIENVENUE sur le site des collectifs d'AC ! - Agir ensemble contre le Ch&ocirc;mage</title>
```

19. <http://apeis.org/>

NONE

20. <http://users.skynet.be/sky74032/>

NONE

21. <http://globenet.org/dal/>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<meta name="author" content="Droit au Logement">
<meta name="keywords" lang="fr" content="droit au logement, droits fondamentaux, droit fondamental, logement, logement décent, logement indigne, logement vacant, logement vide, saturnisme, péril, surpeuplement, hébergement, insalubrité, campement, expulsion, expulsion locative, squatt, mal logés, sans logis, sans domicile, sans abri, état de nécessité, réquisition, mendicité, bivouac, mouvement social, mouvements sociaux, hlm, logement social, social, justice, inégalité, pauvreté, pauvre, exclusion, exclu, précarité, spéculation immobilière, hôtel meublé, militant, locataire">
<meta name="description" content="L'Association Droit Au Logement (DAL) a été créée en 1990, par des familles mal-logées ou sans-logis expulsées à Paris de deux immeubles squattés.">
<meta name="Copyright" Content="Droit au Logement">
<meta name="Identifier-url" Content="http://globenet.org/dal">
22. <http://droitsdevant.ouvaton.org/>
NONE
23. <http://www.gisti.org/>
<meta name="title" lang="fr" content="Groupe d'information et de soutien des immigréés">
<meta name='description' lang='fr' content='Conseils juridiques pour les étrangers. Publications et textes de réflexion sur l'immigration et la liberté de circulation.'>
<meta name="keywords" lang="fr" content="étrangers, immigrés, migrants, immigration, politiques migratoires, nationalité, droits, liberté de circulation">
<meta name='description' lang='en' content='French human rights organization. It protects legal and political rights of foreigners and immigrants and advocates freedom of movement across borders.'>
24. <http://observatoire.samizdat.net/>
<meta name="generator" content="pouet">
<title>Observatoire | Accueil</title>
<meta name="author" content="Jean-Pierre Masse">
<meta name="ROBOTS" content="all, follow, index">
<meta name="DESCRIPTION" content="observatoire du droit des usagers et des institutions sociales">
25. <http://ratp.samizdat.net/>
NONE
26. <http://www.raslfront.org/>
NONE

Nigerian Election Sites

1. <http://www.apgawomen.org/>
NONE
2. <http://www.apgafoundation.org/>
NONE
3. <http://www.afrikontakt.com/alliance/>
<META HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=iso-8859-1">
<META NAME="Generator" CONTENT="NetObjects Fusion 4.0.1 for Windows">
<META NAME="GENERATOR" CONTENT="Mozilla/4.04 [en] (Win95; I) [Netscape]">
<META NAME="Author" CONTENT="Afrikontakt">
4. <http://afenifere.virtualave.net/>
NONE
5. <http://www.muhammadubuhari.com/>
<meta name="description" content="Muhammadu Buhari">
<meta name="keywords" content="Muhammadu Buhari, Buhari, ANPP, 2003, elections, nigeria, presidential candidate">
6. <http://www.buhari2003.org/>
NONE
7. <http://buhariokadigbo.com/>
<meta name="GENERATOR" content="Microsoft FrontPage 4.0">
<meta name="ProgId" content="FrontPage.Editor.Document">
<meta name="Microsoft Theme" content="copy-of-straight-edge 000, default">
<meta name="Microsoft Border" content="tlb, default">
= NO DESCRIPTIVE
8. <http://www.buhari.org/pages/1/index.htm>
NONE
9. <http://www.socialistnigeria.org/>
<meta http-equiv="Content-Language" content="en-gb">
<meta http-equiv="Content-Type" content="text/html; charset=windows-1252">
<meta name="Description" content="Socialist news, policies and Marxist analysis, socialist campaigns, anti-war campaigns, support for workers' struggles">
<meta name="keywords" content="democracy, socialist, socialism, socialists, socialist organisations, socialist news, marxism, marxist, organisation, anti-capitalist, anti-globalisation, marx, leninism, lenin, trotsky, trotskyism, strike, leftwing, party, socialist manifesto, manifesto, politics, election campaign, revolution, militant">
<meta name="GENERATOR" content="Microsoft FrontPage 4.0">
<meta name="ProgId" content="FrontPage.Editor.Document">
10. <http://www.nigeriancp.net/>
<META NAME="keywords"
CONTENT="nigeria, ncp, conscience, party, gani, fawehinmi, lanre, banjo, obasanjo, ibb, abacha, dele, giwa, falana, aborisade, wa heed, alabede, ghana, africa, guardian, newspapers, newspaper, tribune, vanguard, amana, new, nigerian, alliance, democracy, h uman, rights, watch, news, abia, adamawa, akwa, ibom, anambra, bauchi, bayelsa, benue, borno, cross, river, delta, ebonyi, ekiti, en ugu, gombe, imo, jigawa, kaduna, kano, katsina, kebbi, birnin, kogi, kwara, lagos, nassarawa, niger, ogun, osun, oyo, plateau, povert y, eradication, muslim, christian, conflict, tribal, politics, peoples, democratical, pdp, anpp, app, nec, national, electoral, world, comm ision, , governor, federal, government, ethnic, sokoto, taraba, yobe, zamfara, inec, bomb, soyinka, ken, saro, wiwa">
<TITLE>national conscience party of nigeria homepage, ncp</TITLE>
11. <http://www.ndnigeria.com/>
<meta http-equiv=OWNER content="arewa network group, designed by iNetworks Canada"><meta name=GENERATOR content="Microsoft FrontPage 5.0"><meta name=ProgId content="FrontPage.Editor.Document"><title>New Democrats</title><meta name=description content="a new political party in nigeria, the new democrats, bringing vision

and leadership to nigeria"><meta name=keywords content="nigeria, politics, election, nd, new democrats, young democrats, africa, votes, voting, issues">

12. <http://www.nopa.net/>
NONE
13. <http://npgg.freecyberzone.com/>
NONE
14. <http://www.hope2003.org/>
NONE
15. <http://www.ikenwachukwu.com/>
NONE
16. <http://www.jimnwobodo.com/>
NONE
17. <http://www.johnnwodo2003.org/>
NONE
18. <http://www.olusegun-obasanjo.com/>
NONE
19. <http://www.okadigbo4president.com/>
NONE
20. <http://www.peoplesmandateparty.org/>
NONE

APPENDIX 30

Creator-Generated Title Metadata for Nigerian Election .arc files for April 17, 2003

The reliability of creator-generated metadata (or lack thereof) can be intimated in the following sample of title metadata from two .arc files containing a total of 960 HTML and 1073 non-HTML objects. Titles were provided by the .dat files that accompany the .arcs:

Page Title Metadata

Cases where title is missing or an error page title is generated:

40 title field IS NULL; that is, no title tag is available for harvest
 39 Contain "404 (file not found)" in title
 40 Contain "File Not Found" in title
 32 Contain "301 Permanently Moved" in title
 2 Contain "500 Server Error" in title

Cases where the title is particularly opaque or useless:

93 Contain "Untitled Page" in title
 48 Have non-root level pages with "index" in title
 6 Name a page "page n" or "New Page 1"

Cases where the majority of pages share the same generic title:

9 THE CAMPAIGN WEBSITE OF PRESIDENT OLUSEGUN OBASANJO
 9 Vote Senator Chuba Okadigbo for President
 8 Buhari.org
 60 Chief Nwodo
 75/77 Democratic Socialist Movement

Cases where the same title is infelicitously assigned to more than one page:

2 pages entitled "events in Canada" (ndnigeria -- one took referred to an event in Canada, one to an event in Nigeria)
 events in Canada www.ndnigeria.com/events3.htm
 events in Canada www.ndnigeria.com/events2.htm

48 pages entitled "index" in nopa.net website, when url path is more descriptive.
 index www.nopa.net/Useful_Information/speeches.shtml
 index www.nopa.net/Useful_Information/national_symbols.html
 index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt1.html
 index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt4.html
 index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt7.html
 index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt9.html
 index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt11.html
 index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt22.html
 index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt26.html
 index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt28.html
 index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt36.html
 index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt44.html
 index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt48.html
 index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt52.html
 index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt54.html
 index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt58.html

[index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt63.html](http://www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt63.html)
[index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt67.html](http://www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt67.html)
[index www.nopa.net/FAQ/faq.html](http://www.nopa.net/FAQ/faq.html)
[index www.nopa.net/Contact/contact.html](http://www.nopa.net/Contact/contact.html)
[index www.nopa.net/Office_Of_The_SSA/aboutus.htm](http://www.nopa.net/Office_Of_The_SSA/aboutus.htm)
[index www.nopa.net/Office_Of_The_SSA/aboutnopa/strategies.htm](http://www.nopa.net/Office_Of_The_SSA/aboutnopa/strategies.htm)
[index www.nopa.net/NOPA_Activities/nationalrebirth/bauchi/bauchi.htm](http://www.nopa.net/NOPA_Activities/nationalrebirth/bauchi/bauchi.htm)
[index www.nopa.net/NOPA_Activities/nationalrebirth/concldecl.htm](http://www.nopa.net/NOPA_Activities/nationalrebirth/concldecl.htm)
[index www.nopa.net/NOPA_Activities/nationalrebirth/pix3.htm](http://www.nopa.net/NOPA_Activities/nationalrebirth/pix3.htm)
[index www.nopa.net/NOPA_Activities/nationalrebirth/pix5.htm](http://www.nopa.net/NOPA_Activities/nationalrebirth/pix5.htm)
[index www.nopa.net/NOPA_Activities/nationalrebirth/pix4.htm](http://www.nopa.net/NOPA_Activities/nationalrebirth/pix4.htm)
[index www.nopa.net/NOPA_Activities/nationalrebirth/abujarebirthssa.htm](http://www.nopa.net/NOPA_Activities/nationalrebirth/abujarebirthssa.htm)
[index www.nopa.net/NOPA_Activities/nationalrebirth/princdecl.htm](http://www.nopa.net/NOPA_Activities/nationalrebirth/princdecl.htm)
[index www.nopa.net/NOPA_Activities/nationalrebirth/abia/abia.htm](http://www.nopa.net/NOPA_Activities/nationalrebirth/abia/abia.htm)
[index www.nopa.net/Office_Of_The_SSA/aboutnopa/objectives.htm](http://www.nopa.net/Office_Of_The_SSA/aboutnopa/objectives.htm)
[index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt61.html](http://www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt61.html)
[index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt49.html](http://www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt49.html)
[index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt53.html](http://www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt53.html)
[index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt46.html](http://www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt46.html)
[index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt43.html](http://www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt43.html)
[index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt41.html](http://www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt41.html)
[index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt38.html](http://www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt38.html)
[index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt29.html](http://www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt29.html)
[index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt25.html](http://www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt25.html)
[index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt17.html](http://www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt17.html)
[index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt6.html](http://www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt6.html)
[index www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt2.html](http://www.nopa.net/Useful_Information/corruption-act2000/anti-corrupt2.html)
[index www.nopa.net/Useful_Information/anti-corrupt.html](http://www.nopa.net/Useful_Information/anti-corrupt.html)
[index www.nopa.net/Useful_Information/economic-policy.html](http://www.nopa.net/Useful_Information/economic-policy.html)
[index www.nopa.net/Useful_Information/speeches.htm](http://www.nopa.net/Useful_Information/speeches.htm)

- 6 pages entitled "Bios and Speeches" in peoplesmandate.org
- Bio and speeches www.peoplesmandateparty.org/Biograhand%20speeches.html
 - Bio and speeches www.peoplesmandateparty.org/chiefbio.html
 - Bio and speeches www.peoplesmandateparty.org/opinionleader2.html
 - Bio and speeches www.peoplesmandateparty.org/communique.html
 - Bio and speeches www.peoplesmandateparty.org/lukebio.htm
 - Bio and speeches www.peoplesmandateparty.org/edbio.html

APPENDIX 31

The Case of the Purloined Metadata:

<http://perso.magic.fr/nac/index2.html>

The .dat and .arc files for this index page as it evolved reveal some interesting features as well as important caveats for metadata collection. What appears to have happened in late 2000 is a rewrite of the page involving the wholesale borrowing of a javascript from a German sports web site, along with the accompanying <meta> tags. At one point (14 December 2000), the page contains the borrowed metadata in its HTML head. By the date of the next Alexa harvest (3 May 2001) the sports-related descriptive and keyword metadata in the head has been replaced by more appropriate description and keyword values.

1. As archived on 13 June 2000

The homepage contains one auto-generated <meta> tag, and the title in the <head> has been normalized using HTML entities for characters with diacritics. The page's <meta> tag reveals the application used to generate the HTML (Adobe Page Mill) and the platform (Win). These pieces of information would be useful as preservation metadata.

1.1 .dat file extract

```

http://perso.magic.fr:80/nac/index2.html 195.115.16.3 20000613231136 alexa/dat 1107
m text/html
s 200
c 89a8804aee8ff798503e30134ef94b
k d1b545477a7675e1e307250fe971ea95
v 8147275
V 2656515
n 6936
t Vive la Guerre populaire au P&eacute;rou, au N&eacute;pal, et partout ailleurs! Vive le marxisme-l&eacute;ninisme-
mao&iuml;sme!
l perso.magic.fr/nac/news.html
l perso.magic.fr/nac/presanim.htm
l perso.magic.fr/nac/yorum.mp3
i perso.magic.fr/nac/images/Mao1925.jpg
l perso.magic.fr/nac/fs/nacpres.html
i perso.magic.fr/nac/fs/5l.jpg
i perso.magic.fr/nac/images/autonom.jpg
l perso.magic.fr/nac/pcpres.html
l perso.magic.fr/nac/tpml.htm
i perso.magic.fr/nac/images/tpmllogo.gif
i perso.magic.fr/nac/images/horn.jpg
l perso.magic.fr/nac/fs/nepal.html
l perso.magic.fr/nac/iecpres.html
l perso.magic.fr/nac/npci1.html
l perso.magic.fr/nac/PCAFGH.html
l perso.magic.fr/nac/sarbepres.html
i perso.magic.fr/nac/images/sarbedaranlogo.jpg
l perso.magic.fr/nac/mlmrim.html
i perso.magic.fr/nac/images/mrilogo.gif
l perso.magic.fr/nac/raf1.html
i perso.magic.fr/nac/images/raf.jpg
l perso.magic.fr/nac/pcerpres.html
i perso.magic.fr/nac/images/pcerlogo.gif
i perso.magic.fr/nac/images/leninlogo.gif

```

1.2. Partial .arc file extract [through the head in the HTML]

```

http://perso.magic.fr:80/nac/index2.html 195.115.16.3 20000613231136 text/html 6936
HTTP/1.1 200 OK

```

```
Date: Tue, 13 Jun 2000 23:11:36 GMT
Server: MOL-UNIX/AP-FP-PHP-PERL-FCGI-XML/29022000/DED
Last-Modified: Tue, 06 Jun 2000 23:54:30 GMT
ETag: "6e292-1a05-393d8f36"
Accept-Ranges: bytes
Content-Length: 6661
Connection: close
Content-Type: text/html
```

```
<HTML>
<HEAD>
  <META NAME="GENERATOR" CONTENT="Adobe PageMill 3.0 Win">
  <TITLE>Vive la Guerre populaire au P&eacute;rou, au N&eacute;pal, et partout ailleurs! Vive le marxisme-
  l&eacute;ninisme-mao&iuml;smel</TITLE>
</HEAD>
```

2. As archived on 14 December 2000

Note the commented out attribution for the javascript that the page has borrowed. The <charset> element has changed, and a number of <meta> tags have been added or borrowed? The content is surely suspect; what about the creation date for the file?

```
http://perso.magic.fr:80/nac/index2.html 195.115.16.3 200012140048 text/html 17438
HTTP/1.1 200 OK
Date: Thu, 14 Dec 2000 00:48:51 GMT
Server: MOL-UNIX/AP-FP-PHP-PERL-FCGI-XML/29022000/DED
Last-Modified: Mon, 30 Oct 2000 00:12:27 GMT
ETag: "6e286-430b-39fcbceb"
Accept-Ranges: bytes
Content-Length: 17163
Connection: close
Content-Type: text/html
```

```
<snip>
<HTML>
<!-- saved from url=(0041)http://www.sport.de/spart/sk1/ski006.php3 -->
<HEAD>
  <META NAME="GENERATOR" CONTENT="Adobe PageMill 3.0 Win">
  <TITLE>Bienvenue sur le site de Front Social</TITLE>
  <META CONTENT="text/html; charset=windows-1252" HTTP-EQUIV="Content-Type">
  <META CONTENT="no-cache" HTTP-EQUIV="Pragma">
  <META CONTENT="0" HTTP-EQUIV="Expires">
  <META CONTENT="Sport sports Baseball Basketball Beach-Volleyball Bob Boxen Bundesliga Bundesligavereine
  Championsleague DEL DFB DFB-Pokal Eishockey Ergebnisse Europameisterschaft Europapokal Fernsehen Football
  Formel1 Formel3 Fußball Golf Hallenmasters Handball Hockey Inline-Skating Leichtathletik Motorbike Motorrad
  Motorsport Nationalmannschaft NBA NFL NHL Reiten Rodeln Schwimmen Skifahren Skispringen Snowboard Sportarten
  Sportnachrichten Surfen Tennis Tischtennis Turniere Uefa-Cup US Open Vereine Volleyball Wassersport WBA WBC
  WBO Weltmeisterschaft Weltrangliste Wimbledon Fußball Motorsport Radsport Volleyball Sport Eishockey Skisport Boxen
  Handball Leichtathletik Pferdesport Schwimmen" NAME="keywords">
  <META CONTENT="Sport Sportnachrichten Sportvereine Ergebnisse Tabellen Ranglisten Bundesliga DEL Formel 1
  Tennis" NAME="description">
  <META CONTENT="thu, 30 mar 2000 12:00:00 GMT" HTTP-EQUIV="date">
  <SCRIPT language="JavaScript" SRC="sport_fichiers/sidiscrpt.js">
  <SCRIPT language="JavaScript">
<!--
var on = "/ima/pfeil_weiss2.gif";
var off = "/ima/pfeil_weiss.gif";
var drauf = "/ima/pfeil_weiss2n.gif";
var raus = "/ima/pfeil_weissn.gif";
var rueber="/ima/newhomepage97/pfeil_gruen.gif";
var runter="/ima/newhomepage97/pfeil_rot.gif";
var auf="/ima/strich1.gif";
```

```
var unten="/ima/empty.gif";
var drin = "/ima/club/balkenr.gif";
var draus = "/ima/club/balkenw.gif";
</snip>
```

Compare the <meta> tags in the following archived <head> material taken from the Wayback Machine's December 3rd version of the German sports page from which the javascript was derived. The creation date <meta> tag in this page: <meta http-equiv="date" content="thu, 30 mar 2000 12:00:00 GMT"> was copied over wholesale into the index2.html page for Le Front Social. The GIGO factor is apparently at play here.

```
<html>
<head><META HTTP-EQUIV="Pragma" CONTENT="no-cache">
<META HTTP-EQUIV="Expires" CONTENT="0">
<META NAME="keywords" CONTENT="Sport sports Baseball Basketball Beach-Volleyball Bob Boxen Bundesliga
Bundesligavereine Championsleague DEL DFB DFB-Pokal Eishockey Ergebnisse Europameisterschaft Europapokal
Fernsehen Football Formel1 Formel3 Fu&szlig;ball Golf Hallenmasters Handball Hockey Inline-Skating Leichtathletik
Motorbike Motorrad Motorsport Nationalmannschaft NBA NFL NHL Reiten Rodeln Schwimmen Skifahren Skispringen
Snowboard Sportarten Sportnachrichten Surfen Tennis Tischtennis Turniere Uefa-Cup US Open Vereine Volleyball
Wassersport WBA WBC WBO Weltmeisterschaft Weltrangliste Wimbledon Fu&szlig;ball Motorsport Radsport Volleyball
Sport Eishockey Skisport Boxen Handball Leichtathletik Pferdesport Schwimmen">
<META NAME="description" CONTENT="Sport Sportnachrichten Sportvereine Ergebnisse Tabellen Ranglisten
Bundesliga DEL Formel 1 Tennis">
<meta http-equiv="date" content="thu, 30 mar 2000 12:00:00 GMT">
  <title>sport.de - Ski Alpin</title>
<base HREF="http://sport.de/spart/sk1/ski006.php3" target="_top">
<SCRIPT LANGUAGE="JavaScript"
src="http://web.archive.org/web/20001203235200js_/http://sport.de/sidiscrpt.js"></SCRIPT>
<SCRIPT LANGUAGE="JavaScript">
<!--
var on = "/ima/pfeil_weiss2.gif";
var off = "/ima/pfeil_weiss.gif";
var drauf = "/ima/pfeil_weiss2n.gif";
var raus = "/ima/pfeil_weissn.gif";
var rueber="/ima/newhomepage97/pfeil_gruen.gif";
var runter="/ima/newhomepage97/pfeil_rot.gif";
etc.
```


APPENDIX 32

Summary of crawl data of 2003 Nigerian election sites
 using mercator, wget and lynx (for daily availability monitoring)
 for the period 4/17/2003 to 12/14/2003
 Cornell University Library Research Department

Lists of sites to be crawled:

1st list dated 4/15/2003, 21 sites:

- afenifere.virtualave.net
- buhariokadigbo.com
- npgg.freecyberzone.com
- www.afrikontakt.com/alliance
- www.aniagolu.org
- www.apgafoundation.org
- www.apgawomen.org
- www.buhari.org
- www.buhari2003.org
- www.hope2003.org
- www.ikenwachukwu.com
- www.jimnwobodo.com
- www.johnnwodo2003.org
- www.muhammadubuhari.com
- www.ndnigeria.com
- www.nopa.net
- www.okadigbo4president.com
- www.olusegun-obasanjo.com
- www.peoplesmandateparty.org
- www.socialistnigeria.org
- www.nigeriancp.net

2nd list dated 4/28/2003, 12 sites:

- www.abdullahiadamu.com
- www.agagu.com
- www.ajuluforanambragovernor.com
- www.ebeano.org
- www.inecnigeria.org
- www.lafoga.org
- www.mbuhari.com
- www.nigeriafirst.org/elections.shtml
- www.otunbagbengadaniel.org
- www.rimionline.com
- www.sarahjibril4president.org
- www.unongo.com

3rd list dated 4/30/2003, 2 sites:

- www.anppusa.org
- www.buhari-okadigbo.com

4th list dated 4/30/2003, 2 sites:

- www.eueomnigeria.org
- www.unnigeriaelections.org

Total sites: 37

Information on wget crawls:

Date of crawl	#sites	which sites	Anomalies
4/17/2003	21	list #1	none
4/24/2003	21	list #1	3 sites recrawled several hours later: www.aniagolu.org, www.buhari2003.org, www.peoplesmandateparty.org
4/30/2003	12	list #2	1 site recrawled less than an hour later: www.ebeano.org
5/01/2003	4	lists #3-4	none

Thus we have 2 wget crawls of list #1, but only one crawl of lists 2-4.

Information on Mercator crawls:

Date of crawl	#sites	which sites	download failures (# pages and % of total)	robot excluded pages (# pages and % of total)	other anomalies	crawl start time	crawl end time	crawl duration (hrs: mins: secs)
4/17/2003	21	list #1	<ul style="list-style-type: none"> 4 (3.9%) from www.socialistnigeria.org 	<ul style="list-style-type: none"> 8 (72.7%) from npgg.freecyberzone.com 	none	Thu Apr 17 09:58:49 PDT 2003	Thu Apr 17 12:46:03 PDT 2003	2:47:14
4/24/2003	21	list #1	<ul style="list-style-type: none"> 1 (100%) from buhariokadigbo.com 	<ul style="list-style-type: none"> 8 (72.7%) from npgg.freecyberzone.com 	buhariokadigbo.com was recrawled about 3 hours later	Thu Apr 24 12:52:13 PDT 2003	Thu Apr 24 13:35:12 PDT 2003	0:42:59
5/01/2003	37	lists #1-4	none	<ul style="list-style-type: none"> 8 (72.7%) from npgg.freecyberzone.com 1 (100%) from www.ebeano.org 	none	Thu May 01 12:01:49 PDT 2003	Thu May 01 14:02:28 PDT 2003	2:00:39
7/3/2003	37	lists #1-4	none	<ul style="list-style-type: none"> 8 (72.7%) from npgg.freecyberzone.com 1 (100%) from www.ebeano.org 	none	Thu Jul 03 13:30:34 PDT 2003	Thu Jul 03 14:29:31 PDT 2003	0:58:57
7/16/2003	37	lists #1-4	<ul style="list-style-type: none"> 1 (100%) from www.ebeano.org 1 (100%) from www.johnnwodo2003.org 	<ul style="list-style-type: none"> 6 (66.7%) from npgg.freecyberzone.com 	none	Wed Jul 16 11:04:54 PDT 2003	Wed Jul 16 12:10:21 PDT 2003	1:05:27

Date of crawl	#sites	which sites	download failures (# pages and % of total)	robot excluded pages (# pages and % of total)	other anomalies	crawl start time	crawl end time	crawl duration (hrs: mins: secs)
8/1/2003	25	lists #1-4	1 (100%) from www.johnnwodo2003.org <ul style="list-style-type: none"> 1 (100%) from www.okadigbo4president.com 1 (100%) from www.unongo.com 	<ul style="list-style-type: none"> none in those crawled 	job failed before completion leaving 12 sites incomplete or missing (see notes below table)	Fri Aug 01 00:01:01 PDT 2003	Fri Aug 01 00:52:50 PDT 2003	0:51:49 or slightly more
8/15/2003	36	lists #1-4	<ul style="list-style-type: none"> 1 (100%) from www.johnnwodo2003.org 1 (100%) from www.okadigbo4president.com 	<ul style="list-style-type: none"> 6 (66.7%) from npgg.freecyberzone.com 	no data: <ul style="list-style-type: none"> www.annpusa.org incomplete crawl: <ul style="list-style-type: none"> www.hope2003.org www.rimionline.com 	Fri Aug 15 06:05:01 PDT 2003	Fri Aug 15 08:35:19 PDT 2003	2:29:18
9/1/2003	37	lists #1-4	<ul style="list-style-type: none"> 32 (26.4%) from www.ebeano.org 1 (100%) from www.johnnwodo2003.org 1 (100%) from www.okadigbo4president.com 	<ul style="list-style-type: none"> 6 (66.7%) from npgg.freecyberzone.com 3 (2.6%) from www.nigeriancp.net 	none	Mon Sep 01 06:05:01 PDT 2003	Mon Sep 01 09:02:12 PDT 2003	2:57:11 (note: www.ebeano.org took 1:58:21 to crawl)
9/15/2003	36	lists #1-4	<ul style="list-style-type: none"> 1 (100.0%) from www.aniagolu.org 1 (100.0%) from www.buhari.org 9 (2.6%) from www.ebeano.org 1 (100%) from www.okadigbo4president.com 6 (4.0%) from www.socialistnigeria.org 	<ul style="list-style-type: none"> 6 (66.7%) from npgg.freecyberzone.com 3 (2.5%) from www.nigeriancp.net 	no data: <ul style="list-style-type: none"> www.apgawomen.org incomplete crawl: <ul style="list-style-type: none"> www.johnnwodo2003.org 	Mon Sep 15 06:05:01 PDT 2003	Mon Sep 15 11:26:39 PDT 2003	5:21:38 (note: www.ebeano.org took 3:57:54 to crawl)

Date of crawl	#sites	which sites	download failures (# pages and % of total)	robot excluded pages (# pages and % of total)	other anomalies	crawl start time	crawl end time	crawl duration (hrs: mins: secs)
10/1/2003	37	lists #1-4	<ul style="list-style-type: none"> • 1 (100.0%) from www.aniagolu.org • 1 (100.0%) from www.apgafoundation.org • 1 (100.0%) from www.buhari.org • 16 (4.5%) from www.ebeano.org • 1 (100%) from www.johnnwodo2003.org • 1 (100%) from www.lafoga.org 	<ul style="list-style-type: none"> • 6 (66.7%) from npgg.freecyberzone.com • 1 (100%) from www.ikenwachukwu.com • 3 (2.5%) from www.nigeriancp.net • 1 (100%) from www.okadigbo4president.com 	incomplete crawl: <ul style="list-style-type: none"> • www.inecnigeria.org 	Wed Oct 01 06:05:01 PDT 2003	Wed Oct 01 12:26:31 PDT 2003	6:21:30 (note: www.ebeano.org took 5:07:52 to crawl)
10/15/2003	37	lists #1-4	<ul style="list-style-type: none"> • 1 (100.0%) from www.aniagolu.org • 1 (100.0%) from www.apgafoundation.org • 1 (100.0%) from www.buhari.org • 1 (100%) from www.johnnwodo2003.org • 1 (100%) from www.lafoga.org • 1 (100%) from www.okadigbo4president.com 	<ul style="list-style-type: none"> • 6 (66.7%) from npgg.freecyberzone.com • 1 (100%) from www.ikenwachukwu.com • 3 (2.7%) from www.nigeriancp.net 	incomplete crawl: <ul style="list-style-type: none"> • www.buhari-okadigbo.com 	Wed Oct 15 06:05:01 PDT 2003	Wed Oct 15 07:47:58 PDT 2003	1:42:57
11/1/2003	37	lists #1-4	<ul style="list-style-type: none"> • 1 (100.0%) from www.aniagolu.org • 1 (100.0%) from www.buhari.org • 1 (100%) from www.lafoga.org • 1 (100%) from www.okadigbo4president.com 	<ul style="list-style-type: none"> • 6 (66.7%) from npgg.freecyberzone.com • 1 (100.0%) from www.apgafoundation.org • 1 (100%) from www.ebeano.org • 1 (100%) from www.ikenwachukwu.com • 3 (2.8%) from www.nigeriancp.net 	incomplete crawl: <ul style="list-style-type: none"> • www.johnnwodo2003.org 	Sat Nov 01 06:05:01 PST 2003	Sat Nov 01 06:59:35 PST 2003	0:54:34

Date of crawl	#sites	which sites	download failures (# pages and % of total)	robot excluded pages (# pages and % of total)	other anomalies	crawl start time	crawl end time	crawl duration (hrs: mins: secs)
11/15/2003	37	lists #1-4	<ul style="list-style-type: none"> • 1 (100.0%) from www.aniagolu.org • 1 (100.0%) from www.buhari.org • 1 (100.0%) from www.johnnwodo2003.org • 1 (100.0%) from www.lafoga.org • 1 (100.0%) from www.okadigbo4president.com 	<ul style="list-style-type: none"> • 6 (66.7%) from npgg.freecyberzone.com • 1 (100.0%) from www.apgafoundation.org • 1 (100.0%) from www.ebeano.org • 1 (100.0%) from www.ikenwachukwu.com/ • 3 (2.8%) from www.nigeriancp.net 		Sat Nov 15 06:05:01 PST 2003	Sat Nov 15 06:58:09 PST 2003	0:53:08
12/1/2003	37	lists #1-4	<ul style="list-style-type: none"> • 1 (100.0%) from www.aniagolu.org • 1 (100.0%) from www.johnnwodo2003.org • 1 (100.0%) from www.lafoga.org • 20 (18.2%) from www.olusegun-obasanjo.com 	<ul style="list-style-type: none"> • 6 (66.7%) from npgg.freecyberzone.com • 1 (100.0%) from www.apgafoundation.org • 1 (100.0%) from www.ebeano.org • 1 (100.0%) from www.ikenwachukwu.com/ • 3 (2.8%) from www.nigeriancp.net 	incomplete crawl: <ul style="list-style-type: none"> • www.buhari.org • www.okadigbo4president.com 	Mon Dec 01 06:05:01 PST 2003	Mon Dec 01 07:10:15 PST 2003	1:05:14

Crawls not completed on 8/1/2003:
 Incomplete:
 www.buhari-okadigbo.com

No data:
 afenifere.virtualave.net
 buhariokadigbo.com
 npgg.freecyberzone.com
 www.abdullahiadamu.com
 www.afrikontakt.com/alliance
 www.agagu.com
 www.ajuluforanambragovernor.com
 www.aniagolu.org
 www.anppusa.org
 www.apgafoundation.org
 www.apgawomen.org

Daily http server monitoring using lynx (as basis for supplemental downtime data):

- www.aniagolu.org no longer had election content starting sometime between 4/24/03 and 4/28/03, and went down completely on 9/9/03 and remains down as of 12/14/03
- www.johnnwodo2003.org went down 7/16/03 and remains down as of 12/14/03
- www.okadigbo4president.com went down 7/30/03 and remains down as of 12/14/03
- www.buhari.org changed servers on 8/25/03 from Apache 1.3.27 to Netscape-Enterprise 6.0. On 9/7/03 it disappeared and has not returned as of 12/14/03
- www.lafoga.org went down on 9/20/03 and has not returned as of 12/14/03
- www.apgafoundation.org went down on 9/23/03 and came back up on 10/29/03, but without election content. Instead offers sale of various prescription pharmaceuticals.
- www.ikenwachukwu.com went down on 9/26/03 and has not returned as of 12/14/03
- www.hope2003.org went down on 10/4/03 and became crawlable again on 10/13/03. However, the mercator crawl of 10/15/03 indicates that the previous election content was gone as of that crawl. As of 11/5/03, the site displays a message saying "The registration for hope2003.org has expired. If you are the owner of this domain, you can renew it by clicking the button below." On 12/2/2003 the site became unavailable and has not returned as of 12/14/2003.

Summary:

As of 12/14/03, of the 37 original sites, 7 are no longer reachable (i.e. they generate a connection error if you try to view them), while an additional 1 is still active domains, but no longer carry Nigerian election or even Nigerian-related content, for a total of 8 out of 37 (21.6%) where election or party-related content that was once available is no longer available.

APPENDIX 33

Site URL	Server(s) used with dates of inception (starting 4/17/03)	Notes
afenifere.virtualave.net	Apache	
buhariokadigbo.com	Apache 1.3.27,.28(9/7),.29(11/16)	
npgg.freecyberzone.com	Apache 1.3.26	
www.afrikontakt.com	Apache 1.3.20,.27(6/17)	
www.aniagolu.org	Apache 1.3.24	no response starting 9/9-12/14
www.apgafoundation.org	Apache 1.3.27,2.0.40(11/5)	no response 9/23-10/29, different content after 10/30-12/14
www.apgawomen.org	Apache 1.3.27	
www.buhari.org	Apache 1.3.27, Netscape-Enterprise 6.0(8/25-9/6)	no response starting 9/7-12/14
www.buhari2003.org	Apache 1.3.22,.27(6/5)	
www.hope2003.org	Apache 1.3.22,.26(10/13),.29(11/25)	no response 10/3-10/12, 12/2-12/14
www.ikenwachukwu.com	Apache 1.3.26	no response 9/16-12/14
www.jimnwobodo.com	Apache 1.3.27,.29(11/20)	
www.johnnwodo2003.org	Apache 1.3.27	no response 7/16-12/14
www.muhammadubuhari.com	Apache 1.3.27, .28(8/23)	
www.ndnigeria.com	Zeus 3.4 and 4.2 (alternates regularly between them)	
www.nopa.net	Apache 1.3.6	
www.okadigbo4president.com	Apache 1.3.27,Apache(9/19-10/2)	no response 7/30-9/18, 10/3-12/14
www.olusegun-obasanjo.com	Apache 1.3.19,.27(8/15)	
www.peoplesmandateparty.org	Apache 1.3.27	
www.socialistnigeria.org	Apache 1.3.27,Apache(8/31)	
www.nigeriancp.net	Microsoft IIS 5.0	
www.abdullahiadamu.com	Apache 1.3.27	
www.agagu.com	Apache 1.3.27,.28(8/10)	no response 7/31-8/9
www.ajuluforanambragovernor.com	tigershark 3.0.99,.102,.105,.111 at various times	
www.ebeano.org	Apache 1.3.20,.12(7/19),.20(7/21)	no response 7/5-7/18
www.inecnigeria.org	Apache 1.3.27,.29(12/1)	
www.lafoga.org	Apache 1.3.12,.27(7/2)	no response 9/20-12/14
www.mbuhari.com	Apache,2.0.46(6/27),.47(7/15)	
www.nigeriafirst.org_elections	Apache 1.3.22	
www.otunbagbengadaniel.org	Microsoft IIS 5.0	
www.rimionline.com	Zeus 3.4 and 4.2 (alternates regularly between them)	
www.sarahjibril4president.org	Apache 1.3.27	
www.unongo.com	Apache 1.3.27,.28(9/17),.27(9/19)	
www.anppusa.org	Apache 1.3.27,.28(8/8)	
www.buhari-okadigbo.com	Apache 1.3.27	
www.eueomnigeria.org	Microsoft IIS 5.0	

www.unnigeriaelections.org

Apache 1.3.27,.29(11/20)

37

Apache
Microsoft IIS
Zeus
tigershark

31
3
2
1

APPENDIX 34

Processed data and commentary on Mercator crawl data and lynx http server monitoring for arl, curl, asia and Nigerian sites

Richard Entlich

December 15, 2003 (some data revised and corrected, March 8, 2004 and again March 16, 2004)

MIME type data on a total frequency of occurrence basis

All crawls showed the same top four mime types—text/html, image/jpeg, image/gif and application/pdf.—in the same order. Those four types represented 92.7%, 99.2%, 97.8% 97.6% of all mimes for the (Association of Research Libraries, US), CURL (Consortium of University Research Libraries, UK), Asia and Nigeria crawls respectively. In looking at the patterns for these important top four types, it is interesting to note that the greatest similarity is between the CURL sites and the Asia sites, with strikingly similar breakdowns within the four. ARL shows a smaller proportion of text/html objects, but more images and pdfs. Nigerian sites showed an even smaller percentage of text/html objects, with over half the total mime objects being jpegs or gifs, by far the highest proportion of any of the crawls.

The political site crawls had the smallest total numbers of different mime-types represented (32 for 9asia and 16 for 0501Nigeria compared to 116 for 5arl and 43 for 6curl). Overall, that would point to less risk for the political sites in terms of content that might be subject to obsolescence or being orphaned, since most of the mime-types accounting for the higher numbers in the ARL and CURL curls are esoteric types. The political sites had somewhat higher percentages of certain proprietary content, such as Flash animations (.09% and .32% for 9asia and 0501Nigeria respectively, compared to .02% and .04% for ARL and CURL) but there was not clear pattern. For example, looking at Microsoft Word, the political sites were in the same general range (.38% and .07% for 9asia and 0501Nigeria, compared to .36% and .24% for ARL and CURL, respectively). (Note: the % occurrence for Flash in the ARL sites isn't shown in the tables below, since it was not within the top 20 mime types for those crawls).

Based on the following crawls. Note that the number of sites represents how many crawls completed sufficiently to produce a valid mime-contents.000000 file and may be less than the number of total sites being monitored in that category.:

- 5arl (126 sites)
- 6curl (26 sites)
- 9asia (52 sites)
- 20030501Nigeria (36 sites)

Top twenty MIME types found for combined crawls and each crawl separately:

combined crawls 1648963 total objects 125 different MIME types	MIME type	# counted	% of total objects
	text/html	912286	55.32
	image/jpeg	358205	21.72
	image/gif	219982	13.34
	application/pdf	53553	3.25
	text/plain	19135	1.16
	application/zip	19091	1.16
	application/postscript	11813	0.72
	image/tiff	9524	0.58
	application/x-dvi	5885	0.36
	application/msword	5664	0.34
	audio/x-pn-realaudio	4201	0.25

	image/png	4181	0.25
	image/mrsid	3937	0.24
	text/xml	3543	0.21
	text/css	1836	0.11
	image/x-djvu	1617	0.10
	application/x-tex	1541	0.09
	application/unknown	1218	0.07
	application/octet-stream	1138	0.07
	text/sgml	989	0.06
Total		1639339	99.42

crawl 5arl 1390336 total objects 116 different MIME types	MIME type	# counted	% of total objects
	text/html	735279	52.88
	image/jpeg	303599	21.84
	image/gif	198994	14.31
	application/pdf	50505	3.63
	application/zip	19011	1.37
	text/plain	18656	1.34
	application/postscript	11808	0.85
	image/tiff	9523	0.68
	application/x-dvi	5885	0.42
	application/msword	4970	0.36
	image/png	4081	0.29
	image/mrsid	3937	0.28
	audio/x-pn-realaudio	3550	0.26
	text/xml	3390	0.24
	text/css	1659	0.12
	image/x-djvu	1617	0.12
	application/x-tex	1541	0.11
	application/unknown	1191	0.09
	application/octet-stream	1086	0.08
	text/sgml	988	0.07
Total		1381270	99.35

crawl 6curl 190727 total objects 43 different MIME types	MIME type	# counted	% of total objects
	text/html	131528	68.96
	image/jpeg	39783	20.86
	image/gif	15787	8.28
	application/pdf	2162	1.13
	application/msword	451	0.24
	text/plain	316	0.17
	text/css	125	0.07
	audio/x-pn-realaudio	94	0.05
	application/x-shockwave-flash	73	0.04
	text/xml	69	0.04
	application/vnd.ms-powerpoint	48	0.03

	application/x-javascript	39	0.02
	application/octet-stream	37	0.02
	application/unknown	27	0.01
	application/vnd.ms-excel	23	0.01
	application/rtf	23	0.01
	text/rtf	17	0.01
	image/png	17	0.01
	video/mpeg	15	0.01
	audio/mid	13	0.01
Total		190647	99.96

crawl 9asia 62333 total objects 32 different MIME types	MIME type	# counted	% of total objects
	text/html	42990	68.97
	image/jpeg	12523	20.09
	image/gif	4649	7.46
	application/pdf	796	1.28
	audio/x-pn-realaudio	557	0.89
	application/msword	239	0.38
	text/plain	158	0.25
	application/zip	74	0.12
	application/x-shockwave-flash	59	0.09
	text/xml	57	0.09
	image/png	57	0.09
	application/x-javascript	41	0.07
	text/css	36	0.06
	application/x-macbinary	29	0.05
	image/x-png	15	0.02
	video/quicktime	13	0.02
	text/rtf	6	0.01
	image/bmp	5	0.01
	application/x-zip-compressed	5	0.01
	application/vnd.ms-excel	4	0.01
Total		62313	99.97

Nigeria May 1, 2003 5567 total objects 16 different MIME types	MIME type	# counted	% of total objects
	text/html	2489	44.71
	image/jpeg	2300	41.31
	image/gif	552	9.92
	application/pdf	90	1.62
	text/xml	27	0.49
	image/png	26	0.47
	application/x-shockwave-flash	18	0.32
	text/css	16	0.29
	application/x-javascript	14	0.25
	application/octet-stream	12	0.22

	text/rtf	9	0.16
	text/plain	5	0.09
	application/msword	4	0.07
	video/x-ms-wmv	2	0.04
	audio/midi	2	0.04
	audio/mpeg	1	0.02
Total		5567	100.00

Consolidated version of above tables, showing % occurrence for 11 common MIME types (though not the top eleven for each crawl and all combined crawls)

Relative percentage occurrence of commonly occurring MIME types by object					
MIME type	All (N=1648963)	ARL (N=1390336)	CURL (N=190727)	Asia (N=62333)	Nigerian (N=5567)
text/html	55.3	52.9	69.0	69.0	44.7
image/jpeg	21.7	21.8	20.9	20.1	41.3
image/gif	13.3	14.3	8.3	7.5	9.9
application/pdf	3.2	3.6	1.1	1.3	1.6
text/plain	1.2	1.3	0.2	0.3	0.1
application/zip	1.2	1.4	0.0	0.1	0.0
application/postscript	0.7	0.8	0.0	0.0	0.0
image/tiff	0.6	0.7	0.0	0.0	0.0
application/x-dvi	0.4	0.4	0.0	0.0	0.0
application/msword	0.3	0.4	0.2	0.4	0.1
audio/x-pn-realaudio	0.3	0.3	0.0	0.9	0.0

Consolidated version of above tables showing top ten MIME types by frequency of total occurrence for each of the crawls and all combined crawls.

Top ten MIME types by frequency of total objects (most common first)					
Rank	All	ARL	CURL	Asia	Nigeria
1	text/html	text/html	text/html	text/html	text/html
2	image/jpeg	image/jpeg	image/jpeg	image/jpeg	image/jpeg
3	image/gif	image/gif	image/gif	image/gif	image/gif
4	application/pdf	application/pdf	application/pdf	application/pdf	application/pdf
5	text/plain	application/zip	application/msword	audio/x-pn-realaudio	text/xml
6	application/zip	text/plain	text/plain	application/msword	image/png
7	application/postscript	application/postscript	text/css	text/plain	application/x-shockwave-flash
8	image/tiff	image/tiff	audio/x-pn-realaudio	application/zip	text/css
9	application/x-dvi	application/x-dvi	application/x-shockwave-flash	application/x-shockwave-flash	application/x-javascript
10	application/msword	application/msword	text/xml	text/xml	application/octet-stream

MIME type data on a site basis (i.e., a MIME type is counted if there was even a single instance of it on a site)

MIME types that occurred at least once on at least 10 percent of the crawled sites, for all crawls, and each crawl separately

MIME type (all crawls, 240 sites)	#	%
text/html	240	100.0
image/jpeg	226	94.2
image/gif	225	93.8
application/pdf	165	68.8
text/css	146	60.8
application/msword	129	53.8
text/plain	125	52.1
application/x-javascript	117	48.8
application/vnd.ms-powerpoint	72	30.0
text/xml	71	29.6
application/octet-stream	71	29.6
image/png	67	27.9
application/x-shockwave-flash	62	25.8
application/vnd.ms-excel	50	20.8
application/zip	47	19.6
audio/x-pn-realaudio	44	18.3
video/quicktime	39	16.3
text/rtf	38	15.8
audio/x-wav	34	14.2
image/tiff	27	11.3

MIME type (5arl, 126 sites)	#	%
text/html	126	100.0
image/gif	122	96.8
image/jpeg	121	96.0
application/pdf	111	88.1
text/css	103	81.7
application/msword	93	73.8
text/plain	93	73.8
application/x-javascript	90	71.4
application/vnd.ms-powerpoint	64	50.8
application/octet-stream	58	46.0
text/xml	56	44.4
image/png	55	43.7
application/vnd.ms-excel	45	35.7
application/x-shockwave-flash	43	34.1
application/zip	38	30.2
video/quicktime	37	29.4
audio/x-pn-realaudio	36	28.6
audio/x-wav	29	23.0
text/rtf	28	22.2
image/tiff	26	20.6
application/postscript	17	13.5
application/x-msmetafile	17	13.5
video/mpeg	17	13.5
application/mac-binhex40	16	12.7
audio/basic	15	11.9

MIME types (6curl, 26 sites)	#	%
image/gif	26	100.0
text/html	26	100.0
image/jpeg	25	96.2
application/pdf	24	92.3
text/css	21	80.8
application/msword	18	69.2
text/plain	14	53.8
application/x-javascript	10	38.5
application/octet-stream	9	34.6
application/vnd.ms-powerpoint	7	26.9
text/rtf	6	23.1
text/xml	6	23.1
application/x-shockwave-flash	5	19.2
audio/x-wav	4	15.4
image/png	4	15.4
application/rtf	3	11.5
application/vnd.ms-excel	3	11.5
application/x-tar	3	11.5

MIME type (9asia, 52 sites)	#	%
image/gif	52	100.0
text/html	48	92.3
image/jpeg	45	86.5
application/pdf	22	42.3
text/css	17	32.7
application/msword	15	28.8
text/plain	12	23.1
application/x-javascript	9	17.3
application/octet-stream	8	15.4
application/vnd.ms-powerpoint	7	13.5
text/rtf	6	11.5
text/xml	6	11.5
application/x-shockwave-flash	6	11.5

MIME type (0501Nigeria, 36 sites)	#	%
text/html	36	100.0
image/jpeg	32	88.9
image/gif	32	88.9
text/css	10	27.8
application/x-javascript	8	22.2
application/pdf	8	22.2
application/x-shockwave-flash	6	16.7

Consolidated view of above tables, showing rank order and % of sites of top ten MIME types on a per site basis, for all crawls consolidated, and each crawl separately.

Top ten MIME types showing the percentage of all sites within each category that had at least one example of the type										
Rank	All sites (N=240)	%	ARL (N=126)	%	CURL (N=26)	%	Asia (N=52)	%	Nigeria (N=36)	%
1	text/html	100.0	text/html	100.0	image/gif & text/html (tie)	100.0	text/html	100.0	text/html	100.0
2	image/jpeg	94.2	image/gif	96.8	image/jpeg	96.2	image/jpeg	92.3	image/jpeg & image/gif (tie)	88.9
3	image/gif	93.8	image/jpeg	96.0	application/pdf	92.3	image/gif	86.5	text/css	27.8
4	application/pdf	68.8	application/pdf	88.1	text/css	80.8	application/pdf	42.3	application/x-javascript & application/pdf (tie)	22.2
5	text/css	60.8	text/css	81.7	application/msword	69.2	text/plain	32.7	application/x-shockwave-flash	16.7
6	application/msword	53.8	application/msword & text/plain (tie)	73.8	text/plain	53.8	application/msword	28.8	text/xml & application/msword (tie)	8.3
7	text/plain	52.1	application/x-javascript	71.4	application/x-javascript	38.5	text/css	23.1	image/png & audio/midi (tie)	5.6
8	application/x-javascript	48.8	application/vnd.ms-powerpoint	50.8	application/octet-stream	34.6	application/x-javascript	17.3	video/x-ms-wmv, text/rtf, text/plain, audio/mpeg & application/octet-stream (tie)	2.8
9	application/vnd.ms-powerpoint	30.0	application/octet-stream	46.0	application/vnd.ms-powerpoint	26.9	application/x-shockwave-flash	15.4	n/a	
10	application/octet-stream & text/xml (tie)	29.6	text/xml	44.4	text/rtf & text/xml (tie)	23.1	application/zip	13.5	n/a	

Curatorial Investigation

Document Count Data and average numbers of pages/site, mimes/site and mimes/page

Note that the number of sites in each crawl for the page count data as opposed to the mime-type data is different. This is based on whether data was available for each type.

In general, the ARL sites were the largest, followed by CURL, Asia and Nigeria. The political Asia and Nigeria sites were both considerably smaller than either the ARL or CURL sites. However, on the only per page measure given here (mimes/page) they were closely comparable to the much larger sites.

For the mimes/page data, the numbers are based on the same number and group of sites only for the CURL and Nigeria sites. In those cases, the page counts and mime counts are comparable and the figures of 0.82 and 0.89 mimes/page should be accurate. For the Asia and ARL crawls, the page count data is based on a subset of the sites used for the mime count data. Therefore the actual page count for the full set of sites would be higher, driving the average mimes/page even lower. Assuming that the missing page counts were for sites with the same average number of pages as those actually counted, an "adjusted" mimes/page figures is given in the table below.

As to why all the figures for average mimes/page are less than one, I'm not really sure. I do know that mercator tends to overcount pages and it may very well undercount mimes. My expectation is that this number would be at least a little greater than one, since intuitively, one would expect that the average page has a minimum of one mime, and that quite a few have more than one.

Crawl	5arl	6curl	9asia	0501Nigeria
Total pages	1,420,028	231,284	70,241	6238
# sites	107	26	49	36
Avg. pages/site	13271	8896	1433	173
Total mimes	1,390,336	190,727	62,333	5,567
# sites	126	26	52	36
Avg. mimes per site	11034	7336	1199	155
Avg mimes/page based on actual data	0.98	0.82	0.89	0.89
Adjusted avg. mimes/page (based on page count data proportional to the no. of sites for mime count data)	0.83	0.82	0.84	0.89

Document count data for Nigerian sites across all crawls (shows total document count for each crawl)

Note: Document counts of 2 usually mean the site was either unavailable (down, no DNS entry, etc.) or had a robots.txt exclusion for the entire site in place.

Site URL	Date of crawl (yyyymmdd)	# documents reported by mercator
afenifere.virtualave.net	20030417	10
afenifere.virtualave.net	20030424	10
afenifere.virtualave.net	20030501	10
afenifere.virtualave.net	20030703	10
afenifere.virtualave.net	20030716	10
afenifere.virtualave.net	20030815	10
afenifere.virtualave.net	20030901	10
afenifere.virtualave.net	20030915	10
afenifere.virtualave.net	20031001	11
afenifere.virtualave.net	20031015	11

Curatorial Investigation

afenifere.virtualave.net	20031101	11
afenifere.virtualave.net	20031115	11
afenifere.virtualave.net	20031201	11
buhariokadigbo.com	20030417	56
buhariokadigbo.com	20030424	2
buhariokadigbo.com	20030501	56
buhariokadigbo.com	20030703	56
buhariokadigbo.com	20030716	56
buhariokadigbo.com	20030815	56
buhariokadigbo.com	20030901	56
buhariokadigbo.com	20030915	56
buhariokadigbo.com	20031001	56
buhariokadigbo.com	20031015	56
buhariokadigbo.com	20031101	56
buhariokadigbo.com	20031115	56
buhariokadigbo.com	20031201	56
npgg.freecyberzone.com	20030417	12
npgg.freecyberzone.com	20030424	12
npgg.freecyberzone.com	20030501	12
npgg.freecyberzone.com	20030703	10
npgg.freecyberzone.com	20030716	10
npgg.freecyberzone.com	20030815	10
npgg.freecyberzone.com	20030901	10
npgg.freecyberzone.com	20030915	10
npgg.freecyberzone.com	20031001	10
npgg.freecyberzone.com	20031015	10
npgg.freecyberzone.com	20031101	10
npgg.freecyberzone.com	20031115	10
npgg.freecyberzone.com	20031201	10
www.afrikontakt.com	20030417	51
www.afrikontakt.com	20030424	51
www.afrikontakt.com	20030501	51
www.afrikontakt.com	20030703	51
www.afrikontakt.com	20030716	51
www.afrikontakt.com	20030815	51
www.afrikontakt.com	20030901	51
www.afrikontakt.com	20030915	51
www.afrikontakt.com	20031001	51
www.afrikontakt.com	20031015	51
www.afrikontakt.com	20031101	51
www.afrikontakt.com	20031115	51
www.afrikontakt.com	20031201	51
www.aniagolu.org	20030417	39
www.aniagolu.org	20030424	39
www.aniagolu.org	20030501	50
www.aniagolu.org	20030703	68
www.aniagolu.org	20030716	68
www.aniagolu.org	20030815	73
www.aniagolu.org	20030901	73
www.aniagolu.org	20030915	2
www.aniagolu.org	20031001	2
www.aniagolu.org	20031015	2
www.aniagolu.org	20031101	2
www.aniagolu.org	20031115	2
www.aniagolu.org	20031201	2
www.apgafoundation.org	20030417	93
www.apgafoundation.org	20030424	93
www.apgafoundation.org	20030501	93
www.apgafoundation.org	20030703	93

Curatorial Investigation

www.apgafoundation.org	20030716	93
www.apgafoundation.org	20030815	93
www.apgafoundation.org	20030901	93
www.apgafoundation.org	20030915	93
www.apgafoundation.org	20031001	2
www.apgafoundation.org	20031015	2
www.apgafoundation.org	20031101	2
www.apgafoundation.org	20031115	2
www.apgafoundation.org	20031201	2
www.apgawomen.org	20030417	34
www.apgawomen.org	20030424	34
www.apgawomen.org	20030501	34
www.apgawomen.org	20030703	33
www.apgawomen.org	20030716	33
www.apgawomen.org	20030815	33
www.apgawomen.org	20030901	33
www.apgawomen.org	20031001	33
www.apgawomen.org	20031015	33
www.apgawomen.org	20031101	33
www.apgawomen.org	20031115	33
www.apgawomen.org	20031201	32
www.buhari.org	20030417	53
www.buhari.org	20030424	55
www.buhari.org	20030501	58
www.buhari.org	20030703	60
www.buhari.org	20030716	60
www.buhari.org	20030801	60
www.buhari.org	20030815	60
www.buhari.org	20030901	3
www.buhari.org	20030915	2
www.buhari.org	20031001	2
www.buhari.org	20031015	2
www.buhari.org	20031101	2
www.buhari.org	20031115	2
www.buhari2003.org	20030417	141
www.buhari2003.org	20030424	143
www.buhari2003.org	20030501	147
www.buhari2003.org	20030703	151
www.buhari2003.org	20030716	151
www.buhari2003.org	20030801	151
www.buhari2003.org	20030815	151
www.buhari2003.org	20030901	151
www.buhari2003.org	20030915	151
www.buhari2003.org	20031001	151
www.buhari2003.org	20031015	151
www.buhari2003.org	20031101	151
www.buhari2003.org	20031115	151
www.buhari2003.org	20031201	151
www.hope2003.org	20030417	54
www.hope2003.org	20030424	54
www.hope2003.org	20030501	54
www.hope2003.org	20030703	54
www.hope2003.org	20030716	54
www.hope2003.org	20030801	54
www.hope2003.org	20030901	54
www.hope2003.org	20030915	54
www.hope2003.org	20031001	54
www.hope2003.org	20031015	2
www.hope2003.org	20031101	2
www.hope2003.org	20031115	2

Curatorial Investigation

www.hope2003.org	20031201	2
www.ikenwachukwu.com	20030417	20
www.ikenwachukwu.com	20030424	20
www.ikenwachukwu.com	20030501	20
www.ikenwachukwu.com	20030703	20
www.ikenwachukwu.com	20030716	20
www.ikenwachukwu.com	20030801	20
www.ikenwachukwu.com	20030815	20
www.ikenwachukwu.com	20030901	20
www.ikenwachukwu.com	20030915	20
www.ikenwachukwu.com	20031001	2
www.ikenwachukwu.com	20031015	2
www.ikenwachukwu.com	20031101	2
www.ikenwachukwu.com	20031115	2
www.ikenwachukwu.com	20031201	2
www.jimnwobodo.com	20030417	39
www.jimnwobodo.com	20030424	39
www.jimnwobodo.com	20030501	39
www.jimnwobodo.com	20030703	39
www.jimnwobodo.com	20030716	39
www.jimnwobodo.com	20030801	39
www.jimnwobodo.com	20030815	39
www.jimnwobodo.com	20030901	39
www.jimnwobodo.com	20030915	39
www.jimnwobodo.com	20031001	39
www.jimnwobodo.com	20031015	39
www.jimnwobodo.com	20031101	39
www.jimnwobodo.com	20031115	39
www.jimnwobodo.com	20031201	39
www.johnnwodo2003.org	20030417	91
www.johnnwodo2003.org	20030424	91
www.johnnwodo2003.org	20030501	91
www.johnnwodo2003.org	20030703	91
www.johnnwodo2003.org	20030716	2
www.johnnwodo2003.org	20030801	2
www.johnnwodo2003.org	20030815	2
www.johnnwodo2003.org	20030901	2
www.johnnwodo2003.org	20031001	2
www.johnnwodo2003.org	20031015	2
www.johnnwodo2003.org	20031115	2
www.johnnwodo2003.org	20031201	2
www.muhammadubuhari.com	20030417	27
www.muhammadubuhari.com	20030424	27
www.muhammadubuhari.com	20030501	28
www.muhammadubuhari.com	20030703	30
www.muhammadubuhari.com	20030716	30
www.muhammadubuhari.com	20030801	30
www.muhammadubuhari.com	20030815	30
www.muhammadubuhari.com	20030901	30
www.muhammadubuhari.com	20030915	30
www.muhammadubuhari.com	20031001	30
www.muhammadubuhari.com	20031015	30
www.muhammadubuhari.com	20031101	30
www.muhammadubuhari.com	20031115	30
www.muhammadubuhari.com	20031201	30
www.ndnigeria.com	20030417	572
www.ndnigeria.com	20030424	572
www.ndnigeria.com	20030501	572
www.ndnigeria.com	20030703	572

Curatorial Investigation

www.ndnigeria.com	20030716	572
www.ndnigeria.com	20030801	572
www.ndnigeria.com	20030815	572
www.ndnigeria.com	20030901	572
www.ndnigeria.com	20030915	576
www.ndnigeria.com	20031001	577
www.ndnigeria.com	20031015	580
www.ndnigeria.com	20031101	580
www.ndnigeria.com	20031115	580
www.ndnigeria.com	20031201	585
www.nopa.net	20030417	705
www.nopa.net	20030424	717
www.nopa.net	20030501	717
www.nopa.net	20030703	719
www.nopa.net	20030716	719
www.nopa.net	20030801	719
www.nopa.net	20030815	719
www.nopa.net	20030901	719
www.nopa.net	20030915	719
www.nopa.net	20031001	719
www.nopa.net	20031015	719
www.nopa.net	20031101	719
www.nopa.net	20031115	719
www.nopa.net	20031201	719
www.okadigbo4president.com	20030417	33
www.okadigbo4president.com	20030424	33
www.okadigbo4president.com	20030501	33
www.okadigbo4president.com	20030703	33
www.okadigbo4president.com	20030716	33
www.okadigbo4president.com	20030801	2
www.okadigbo4president.com	20030815	2
www.okadigbo4president.com	20030901	2
www.okadigbo4president.com	20030915	2
www.okadigbo4president.com	20031001	2
www.okadigbo4president.com	20031015	2
www.okadigbo4president.com	20031101	2
www.okadigbo4president.com	20031115	2
www.olusegun-obasanjo.com	20030417	44
www.olusegun-obasanjo.com	20030424	46
www.olusegun-obasanjo.com	20030501	46
www.olusegun-obasanjo.com	20030703	119
www.olusegun-obasanjo.com	20030716	119
www.olusegun-obasanjo.com	20030801	120
www.olusegun-obasanjo.com	20030815	133
www.olusegun-obasanjo.com	20030901	133
www.olusegun-obasanjo.com	20030915	133
www.olusegun-obasanjo.com	20031001	133
www.olusegun-obasanjo.com	20031015	133
www.olusegun-obasanjo.com	20031101	133
www.olusegun-obasanjo.com	20031115	133
www.olusegun-obasanjo.com	20031201	111
www.peoplesmandateparty.org	20030417	93
www.peoplesmandateparty.org	20030424	93
www.peoplesmandateparty.org	20030501	93
www.peoplesmandateparty.org	20030703	93
www.peoplesmandateparty.org	20030716	93
www.peoplesmandateparty.org	20030801	93
www.peoplesmandateparty.org	20030815	93
www.peoplesmandateparty.org	20030901	93
www.peoplesmandateparty.org	20030915	93

Curatorial Investigation

www.peoplesmandateparty.org	20031001	93
www.peoplesmandateparty.org	20031015	93
www.peoplesmandateparty.org	20031101	93
www.peoplesmandateparty.org	20031115	93
www.peoplesmandateparty.org	20031201	93
www.socialistnigeria.org	20030417	103
www.socialistnigeria.org	20030424	104
www.socialistnigeria.org	20030501	106
www.socialistnigeria.org	20030703	152
www.socialistnigeria.org	20030716	154
www.socialistnigeria.org	20030801	156
www.socialistnigeria.org	20030815	156
www.socialistnigeria.org	20030901	156
www.socialistnigeria.org	20030915	152
www.socialistnigeria.org	20031001	185
www.socialistnigeria.org	20031015	186
www.socialistnigeria.org	20031101	189
www.socialistnigeria.org	20031115	190
www.socialistnigeria.org	20031201	205
www.nigeriancp.net	20030417	102
www.nigeriancp.net	20030424	102
www.nigeriancp.net	20030501	100
www.nigeriancp.net	20030703	89
www.nigeriancp.net	20030716	106
www.nigeriancp.net	20030801	106
www.nigeriancp.net	20030815	107
www.nigeriancp.net	20030901	118
www.nigeriancp.net	20030915	119
www.nigeriancp.net	20031001	119
www.nigeriancp.net	20031015	112
www.nigeriancp.net	20031101	109
www.nigeriancp.net	20031115	109
www.nigeriancp.net	20031201	109
www.abdullahiadamu.com	20030501	2292
www.abdullahiadamu.com	20030703	2288
www.abdullahiadamu.com	20030716	2296
www.abdullahiadamu.com	20030815	2300
www.abdullahiadamu.com	20030901	2324
www.abdullahiadamu.com	20030915	2335
www.abdullahiadamu.com	20031001	2338
www.abdullahiadamu.com	20031015	2343
www.abdullahiadamu.com	20031101	2345
www.abdullahiadamu.com	20031115	2345
www.abdullahiadamu.com	20031201	2345
www.agagu.com	20030501	31
www.agagu.com	20030703	35
www.agagu.com	20030716	35
www.agagu.com	20030815	34
www.agagu.com	20030901	34
www.agagu.com	20030915	34
www.agagu.com	20031001	34
www.agagu.com	20031015	34
www.agagu.com	20031101	34
www.agagu.com	20031115	34
www.agagu.com	20031201	34
www.ajuluforanambragovernor.com	20030501	11
www.ajuluforanambragovernor.com	20030703	11
www.ajuluforanambragovernor.com	20030716	11
www.ajuluforanambragovernor.com	20030815	11

Curatorial Investigation

www.ajuluforanambragovernor.com	20030901	11
www.ajuluforanambragovernor.com	20030915	11
www.ajuluforanambragovernor.com	20031001	11
www.ajuluforanambragovernor.com	20031015	11
www.ajuluforanambragovernor.com	20031101	11
www.ajuluforanambragovernor.com	20031115	11
www.ajuluforanambragovernor.com	20031201	11
www.ebeano.org	20030501	2
www.ebeano.org	20030512	2
www.ebeano.org	20030703	2
www.ebeano.org	20030716	2
www.ebeano.org	20030801	89
www.ebeano.org	20030815	173
www.ebeano.org	20030901	122
www.ebeano.org	20030915	350
www.ebeano.org	20031001	356
www.ebeano.org	20031015	306
www.ebeano.org	20031101	2
www.ebeano.org	20031115	2
www.ebeano.org	20031201	2
www.inecnigeria.org	20030501	338
www.inecnigeria.org	20030703	377
www.inecnigeria.org	20030716	377
www.inecnigeria.org	20030801	380
www.inecnigeria.org	20030815	398
www.inecnigeria.org	20030901	406
www.inecnigeria.org	20030915	406
www.inecnigeria.org	20031015	438
www.inecnigeria.org	20031101	438
www.inecnigeria.org	20031115	438
www.inecnigeria.org	20031201	436
www.lafoga.org	20030501	49
www.lafoga.org	20030703	49
www.lafoga.org	20030716	49
www.lafoga.org	20030801	49
www.lafoga.org	20030815	49
www.lafoga.org	20030901	49
www.lafoga.org	20030915	49
www.lafoga.org	20031001	2
www.lafoga.org	20031015	2
www.lafoga.org	20031101	2
www.lafoga.org	20031115	2
www.lafoga.org	20031201	2
www.mbuhari.com	20030501	114
www.mbuhari.com	20030703	84
www.mbuhari.com	20030716	84
www.mbuhari.com	20030801	84
www.mbuhari.com	20030815	84
www.mbuhari.com	20030901	84
www.mbuhari.com	20030915	84
www.mbuhari.com	20031001	84
www.mbuhari.com	20031015	84
www.mbuhari.com	20031101	84
www.mbuhari.com	20031115	84
www.mbuhari.com	20031201	84
www.otunbagbengadaniel.org	20030501	29
www.otunbagbengadaniel.org	20030703	29
www.otunbagbengadaniel.org	20030716	29

Curatorial Investigation

www.otunbagbengadaniel.org	20030801	29
www.otunbagbengadaniel.org	20030815	29
www.otunbagbengadaniel.org	20030901	29
www.otunbagbengadaniel.org	20030915	29
www.otunbagbengadaniel.org	20031001	29
www.otunbagbengadaniel.org	20031015	29
www.otunbagbengadaniel.org	20031101	29
www.otunbagbengadaniel.org	20031115	29
www.otunbagbengadaniel.org	20031201	29
www.rimionline.com	20030501	41
www.rimionline.com	20030703	41
www.rimionline.com	20030716	41
www.rimionline.com	20030801	41
www.rimionline.com	20030901	41
www.rimionline.com	20030915	41
www.rimionline.com	20031001	41
www.rimionline.com	20031015	51
www.rimionline.com	20031101	51
www.rimionline.com	20031115	51
www.rimionline.com	20031201	51
www.sarahjibril4president.org	20030501	41
www.sarahjibril4president.org	20030703	41
www.sarahjibril4president.org	20030716	41
www.sarahjibril4president.org	20030801	41
www.sarahjibril4president.org	20030815	41
www.sarahjibril4president.org	20030901	41
www.sarahjibril4president.org	20030915	41
www.sarahjibril4president.org	20031001	41
www.sarahjibril4president.org	20031015	41
www.sarahjibril4president.org	20031101	41
www.sarahjibril4president.org	20031115	41
www.sarahjibril4president.org	20031201	41
www.unongo.com	20030501	6
www.unongo.com	20030703	6
www.unongo.com	20030716	6
www.unongo.com	20030801	2
www.unongo.com	20030815	6
www.unongo.com	20030901	6
www.unongo.com	20030915	6
www.unongo.com	20031001	6
www.unongo.com	20031015	6
www.unongo.com	20031101	6
www.unongo.com	20031115	6
www.unongo.com	20031201	6
www.anppusa.org	20030501	76
www.anppusa.org	20030703	10
www.anppusa.org	20030716	76
www.anppusa.org	20030901	76
www.anppusa.org	20030915	76
www.anppusa.org	20031001	76
www.anppusa.org	20031015	76
www.anppusa.org	20031101	76
www.anppusa.org	20031115	76
www.anppusa.org	20031201	76
www.buhari-okadigbo.com	20030501	38
www.buhari-okadigbo.com	20030703	38
www.buhari-okadigbo.com	20030716	38
www.buhari-okadigbo.com	20030815	38
www.buhari-okadigbo.com	20030901	38

Curatorial Investigation

www.buhari-okadigbo.com	20030915	38
www.buhari-okadigbo.com	20031001	38
www.buhari-okadigbo.com	20031101	38
www.buhari-okadigbo.com	20031115	38
www.buhari-okadigbo.com	20031201	38
www.eueomnigeria.org	20030501	95
www.eueomnigeria.org	20030703	102
www.eueomnigeria.org	20030716	102
www.eueomnigeria.org	20030801	102
www.eueomnigeria.org	20030815	102
www.eueomnigeria.org	20030901	102
www.eueomnigeria.org	20030915	102
www.eueomnigeria.org	20031001	102
www.eueomnigeria.org	20031015	102
www.eueomnigeria.org	20031101	102
www.eueomnigeria.org	20031115	102
www.eueomnigeria.org	20031201	102
www.unnigeriaelections.org	20030501	148
www.unnigeriaelections.org	20030703	169
www.unnigeriaelections.org	20030716	169
www.unnigeriaelections.org	20030801	169
www.unnigeriaelections.org	20030815	169
www.unnigeriaelections.org	20030901	169
www.unnigeriaelections.org	20030915	173
www.unnigeriaelections.org	20031001	174
www.unnigeriaelections.org	20031015	174
www.unnigeriaelections.org	20031101	174
www.unnigeriaelections.org	20031115	174
www.unnigeriaelections.org	20031201	179

Byte count data for Nigerian sites across all crawls (shows total byte count for each crawl)

Site URL	Date of crawl (yyyymmdd)	# total bytes reported by mercator
afenifere.virtualave.net	20030417	359644
afenifere.virtualave.net	20030424	358652
afenifere.virtualave.net	20030501	357089
afenifere.virtualave.net	20030703	362202
afenifere.virtualave.net	20030716	362211
afenifere.virtualave.net	20030815	357726
afenifere.virtualave.net	20030901	356796
afenifere.virtualave.net	20030915	353348
afenifere.virtualave.net	20031001	343727
afenifere.virtualave.net	20031015	343727
afenifere.virtualave.net	20031101	347687
afenifere.virtualave.net	20031115	348137
afenifere.virtualave.net	20031201	348137
buhariokadigbo.com	20030417	597537
buhariokadigbo.com	20030424	0
buhariokadigbo.com	20030501	597537
buhariokadigbo.com	20030703	597537
buhariokadigbo.com	20030716	597537
buhariokadigbo.com	20030815	597537
buhariokadigbo.com	20030901	597537
buhariokadigbo.com	20030915	597537
buhariokadigbo.com	20031001	597537
buhariokadigbo.com	20031015	597537

Curatorial Investigation

buhariokadigbo.com	20031101	597537
buhariokadigbo.com	20031115	597537
buhariokadigbo.com	20031201	597537
npgg.freecyberzone.com	20030417	41562
npgg.freecyberzone.com	20030424	41562
npgg.freecyberzone.com	20030501	41456
npgg.freecyberzone.com	20030703	41673
npgg.freecyberzone.com	20030716	41658
npgg.freecyberzone.com	20030815	41659
npgg.freecyberzone.com	20030901	41661
npgg.freecyberzone.com	20030915	42470
npgg.freecyberzone.com	20031001	42470
npgg.freecyberzone.com	20031015	42470
npgg.freecyberzone.com	20031101	42470
npgg.freecyberzone.com	20031115	42470
npgg.freecyberzone.com	20031201	42466
www.afrikontakt.com	20030417	301318
www.afrikontakt.com	20030424	301318
www.afrikontakt.com	20030501	301318
www.afrikontakt.com	20030703	301318
www.afrikontakt.com	20030716	301318
www.afrikontakt.com	20030815	301318
www.afrikontakt.com	20030901	301318
www.afrikontakt.com	20030915	301318
www.afrikontakt.com	20031001	301318
www.afrikontakt.com	20031015	301318
www.afrikontakt.com	20031101	301318
www.afrikontakt.com	20031115	301318
www.afrikontakt.com	20031201	301318
www.aniagolu.org	20030417	1047119
www.aniagolu.org	20030424	1047119
www.aniagolu.org	20030501	972425
www.aniagolu.org	20030703	1211417
www.aniagolu.org	20030716	1313127
www.aniagolu.org	20030815	1453934
www.aniagolu.org	20030901	1459178
www.aniagolu.org	20030915	0
www.aniagolu.org	20031001	0
www.aniagolu.org	20031015	0
www.aniagolu.org	20031101	0
www.aniagolu.org	20031115	0
www.aniagolu.org	20031201	0
www.apgafoundation.org	20030417	1819860
www.apgafoundation.org	20030424	1819860
www.apgafoundation.org	20030501	1819860
www.apgafoundation.org	20030703	1819860
www.apgafoundation.org	20030716	1819860
www.apgafoundation.org	20030815	1819860
www.apgafoundation.org	20030901	1819860
www.apgafoundation.org	20030915	1819860
www.apgafoundation.org	20031001	0
www.apgafoundation.org	20031015	0
www.apgafoundation.org	20031101	1853
www.apgafoundation.org	20031115	1853
www.apgafoundation.org	20031201	1853
www.apgawomen.org	20030417	286349
www.apgawomen.org	20030424	286349
www.apgawomen.org	20030501	286349
www.apgawomen.org	20030703	306750

Curatorial Investigation

www.apgawomen.org	20030716	306750
www.apgawomen.org	20030815	306750
www.apgawomen.org	20030901	306750
www.apgawomen.org	20031001	306750
www.apgawomen.org	20031015	306750
www.apgawomen.org	20031101	306750
www.apgawomen.org	20031115	306750
www.apgawomen.org	20031201	306464
www.buhari.org	20030417	1352212
www.buhari.org	20030424	1397733
www.buhari.org	20030501	1549215
www.buhari.org	20030703	2730127
www.buhari.org	20030716	2730127
www.buhari.org	20030801	2730127
www.buhari.org	20030815	2730127
www.buhari.org	20030901	24909
www.buhari.org	20030915	0
www.buhari.org	20031001	0
www.buhari.org	20031015	0
www.buhari.org	20031101	0
www.buhari.org	20031115	0
www.buhari2003.org	20030417	2990205
www.buhari2003.org	20030424	3026391
www.buhari2003.org	20030501	3569616
www.buhari2003.org	20030703	6155560
www.buhari2003.org	20030716	6155560
www.buhari2003.org	20030801	6155560
www.buhari2003.org	20030815	6155586
www.buhari2003.org	20030901	6155586
www.buhari2003.org	20030915	6155586
www.buhari2003.org	20031001	6155586
www.buhari2003.org	20031015	6155586
www.buhari2003.org	20031101	6155584
www.buhari2003.org	20031115	6155584
www.buhari2003.org	20031201	6155586
www.hope2003.org	20030417	672193
www.hope2003.org	20030424	672241
www.hope2003.org	20030501	672219
www.hope2003.org	20030703	672287
www.hope2003.org	20030716	672287
www.hope2003.org	20030801	672287
www.hope2003.org	20030901	672261
www.hope2003.org	20030915	672223
www.hope2003.org	20031001	672223
www.hope2003.org	20031015	74
www.hope2003.org	20031101	74
www.hope2003.org	20031115	74
www.hope2003.org	20031201	74
www.ikenwachukwu.com	20030417	268521
www.ikenwachukwu.com	20030424	268521
www.ikenwachukwu.com	20030501	268521
www.ikenwachukwu.com	20030703	268521
www.ikenwachukwu.com	20030716	268521
www.ikenwachukwu.com	20030801	268521
www.ikenwachukwu.com	20030815	268521
www.ikenwachukwu.com	20030901	268521
www.ikenwachukwu.com	20030915	268521
www.ikenwachukwu.com	20031001	0
www.ikenwachukwu.com	20031015	0
www.ikenwachukwu.com	20031101	0

Curatorial Investigation

www.ikenwachukwu.com	20031115	0
www.ikenwachukwu.com	20031201	0
www.jimnwobodo.com	20030417	260431
www.jimnwobodo.com	20030424	260431
www.jimnwobodo.com	20030501	260431
www.jimnwobodo.com	20030703	260431
www.jimnwobodo.com	20030716	260431
www.jimnwobodo.com	20030801	260431
www.jimnwobodo.com	20030815	260431
www.jimnwobodo.com	20030901	260431
www.jimnwobodo.com	20030915	260431
www.jimnwobodo.com	20031001	260431
www.jimnwobodo.com	20031015	260431
www.jimnwobodo.com	20031101	260431
www.jimnwobodo.com	20031115	260431
www.jimnwobodo.com	20031201	260431
www.johnnwodo2003.org	20030417	1224385
www.johnnwodo2003.org	20030424	1224385
www.johnnwodo2003.org	20030501	1224385
www.johnnwodo2003.org	20030703	1224385
www.johnnwodo2003.org	20030716	0
www.johnnwodo2003.org	20030801	0
www.johnnwodo2003.org	20030815	0
www.johnnwodo2003.org	20030901	0
www.johnnwodo2003.org	20031001	0
www.johnnwodo2003.org	20031015	0
www.johnnwodo2003.org	20031115	0
www.johnnwodo2003.org	20031201	0
www.muhammadubuhari.com	20030417	819344
www.muhammadubuhari.com	20030424	819344
www.muhammadubuhari.com	20030501	901024
www.muhammadubuhari.com	20030703	2428708
www.muhammadubuhari.com	20030716	2428708
www.muhammadubuhari.com	20030801	2428708
www.muhammadubuhari.com	20030815	2428708
www.muhammadubuhari.com	20030901	2428693
www.muhammadubuhari.com	20030915	2428742
www.muhammadubuhari.com	20031001	2428729
www.muhammadubuhari.com	20031015	2428729
www.muhammadubuhari.com	20031101	2428729
www.muhammadubuhari.com	20031115	2428729
www.muhammadubuhari.com	20031201	2428729
www.ndnigeria.com	20030417	23737270
www.ndnigeria.com	20030424	23736471
www.ndnigeria.com	20030501	23737270
www.ndnigeria.com	20030703	23749206
www.ndnigeria.com	20030716	23749206
www.ndnigeria.com	20030801	23748185
www.ndnigeria.com	20030815	23748545
www.ndnigeria.com	20030901	23748545
www.ndnigeria.com	20030915	23764608
www.ndnigeria.com	20031001	23766942
www.ndnigeria.com	20031015	23777595
www.ndnigeria.com	20031101	23777595
www.ndnigeria.com	20031115	23777595
www.ndnigeria.com	20031201	23809976
www.nopa.net	20030417	9578469
www.nopa.net	20030424	9961375
www.nopa.net	20030501	9961375

Curatorial Investigation

www.nopa.net	20030703	9986399
www.nopa.net	20030716	9986399
www.nopa.net	20030801	9986399
www.nopa.net	20030815	9986399
www.nopa.net	20030901	9986399
www.nopa.net	20030915	9986399
www.nopa.net	20031001	9986399
www.nopa.net	20031015	9986399
www.nopa.net	20031101	9986399
www.nopa.net	20031115	9986399
www.nopa.net	20031201	9986399
www.okadigbo4president.com	20030417	144423
www.okadigbo4president.com	20030424	144423
www.okadigbo4president.com	20030501	144423
www.okadigbo4president.com	20030703	144423
www.okadigbo4president.com	20030716	144423
www.okadigbo4president.com	20030801	0
www.okadigbo4president.com	20030815	0
www.okadigbo4president.com	20030901	0
www.okadigbo4president.com	20030915	0
www.okadigbo4president.com	20031001	26
www.okadigbo4president.com	20031015	0
www.okadigbo4president.com	20031101	0
www.okadigbo4president.com	20031115	0
www.olusegun-obasanjo.com	20030417	818020
www.olusegun-obasanjo.com	20030424	1132591
www.olusegun-obasanjo.com	20030501	1210758
www.olusegun-obasanjo.com	20030703	1732859
www.olusegun-obasanjo.com	20030716	1738484
www.olusegun-obasanjo.com	20030801	2970554
www.olusegun-obasanjo.com	20030815	2832733
www.olusegun-obasanjo.com	20030901	2832733
www.olusegun-obasanjo.com	20030915	2832733
www.olusegun-obasanjo.com	20031001	2832733
www.olusegun-obasanjo.com	20031015	2832733
www.olusegun-obasanjo.com	20031101	2832733
www.olusegun-obasanjo.com	20031115	2832733
www.olusegun-obasanjo.com	20031201	1118616
www.peoplesmandateparty.org	20030417	1400843
www.peoplesmandateparty.org	20030424	1400843
www.peoplesmandateparty.org	20030501	1400843
www.peoplesmandateparty.org	20030703	1400843
www.peoplesmandateparty.org	20030716	1400843
www.peoplesmandateparty.org	20030801	1400843
www.peoplesmandateparty.org	20030815	1400843
www.peoplesmandateparty.org	20030901	1400843
www.peoplesmandateparty.org	20030915	1400843
www.peoplesmandateparty.org	20031001	1400843
www.peoplesmandateparty.org	20031015	1400843
www.peoplesmandateparty.org	20031101	1400843
www.peoplesmandateparty.org	20031115	1400843
www.peoplesmandateparty.org	20031201	1400843
www.socialistnigeria.org	20030417	2351693
www.socialistnigeria.org	20030424	2861098
www.socialistnigeria.org	20030501	2880814
www.socialistnigeria.org	20030703	3575710
www.socialistnigeria.org	20030716	3605389
www.socialistnigeria.org	20030801	3673179
www.socialistnigeria.org	20030815	3673179
www.socialistnigeria.org	20030901	3673179

Curatorial Investigation

www.socialistnigeria.org	20030915	3362643
www.socialistnigeria.org	20031001	4230639
www.socialistnigeria.org	20031015	4251239
www.socialistnigeria.org	20031101	4307111
www.socialistnigeria.org	20031115	4300132
www.socialistnigeria.org	20031201	4762411
www.nigeriancp.net	20030417	2581755
www.nigeriancp.net	20030424	2581755
www.nigeriancp.net	20030501	2575060
www.nigeriancp.net	20030703	2303224
www.nigeriancp.net	20030716	2864230
www.nigeriancp.net	20030801	2863946
www.nigeriancp.net	20030815	2891211
www.nigeriancp.net	20030901	2924286
www.nigeriancp.net	20030915	2919555
www.nigeriancp.net	20031001	2924404
www.nigeriancp.net	20031015	2450968
www.nigeriancp.net	20031101	2452802
www.nigeriancp.net	20031115	2452802
www.nigeriancp.net	20031201	2452802
www.abdullahiadamu.com	20030501	40220442
www.abdullahiadamu.com	20030703	39696141
www.abdullahiadamu.com	20030716	39802206
www.abdullahiadamu.com	20030815	39844794
www.abdullahiadamu.com	20030901	40122169
www.abdullahiadamu.com	20030915	40267879
www.abdullahiadamu.com	20031001	40308802
www.abdullahiadamu.com	20031015	40378491
www.abdullahiadamu.com	20031101	40407035
www.abdullahiadamu.com	20031115	40407035
www.abdullahiadamu.com	20031201	40407035
www.agagu.com	20030501	538226
www.agagu.com	20030703	559757
www.agagu.com	20030716	559757
www.agagu.com	20030815	388085
www.agagu.com	20030901	388085
www.agagu.com	20030915	388085
www.agagu.com	20031001	388085
www.agagu.com	20031015	388085
www.agagu.com	20031101	388085
www.agagu.com	20031115	388085
www.agagu.com	20031201	388085
www.ajuluforanambragovernor.com	20030501	102093
www.ajuluforanambragovernor.com	20030703	102094
www.ajuluforanambragovernor.com	20030716	102094
www.ajuluforanambragovernor.com	20030815	102094
www.ajuluforanambragovernor.com	20030901	102094
www.ajuluforanambragovernor.com	20030915	102094
www.ajuluforanambragovernor.com	20031001	102094
www.ajuluforanambragovernor.com	20031015	102094
www.ajuluforanambragovernor.com	20031101	102094
www.ajuluforanambragovernor.com	20031115	102094
www.ajuluforanambragovernor.com	20031201	102094
www.ebeano.org	20030501	0
www.ebeano.org	20030512	0
www.ebeano.org	20030703	0
www.ebeano.org	20030716	0
www.ebeano.org	20030801	1244816
www.ebeano.org	20030815	2192514

Curatorial Investigation

www.ebeano.org	20030901	1070302
www.ebeano.org	20030915	4294678
www.ebeano.org	20031001	4083432
www.ebeano.org	20031015	3659089
www.ebeano.org	20031101	0
www.ebeano.org	20031115	0
www.ebeano.org	20031201	0
www.inecnigeria.org	20030501	4992331
www.inecnigeria.org	20030703	5824674
www.inecnigeria.org	20030716	5824674
www.inecnigeria.org	20030801	5893442
www.inecnigeria.org	20030815	6302920
www.inecnigeria.org	20030901	6483473
www.inecnigeria.org	20030915	6483473
www.inecnigeria.org	20031015	7043337
www.inecnigeria.org	20031101	7043337
www.inecnigeria.org	20031115	7043337
www.inecnigeria.org	20031201	7097247
www.lafoga.org	20030501	776029
www.lafoga.org	20030703	776027
www.lafoga.org	20030716	776027
www.lafoga.org	20030801	776027
www.lafoga.org	20030815	776027
www.lafoga.org	20030901	776027
www.lafoga.org	20030915	776027
www.lafoga.org	20031001	0
www.lafoga.org	20031015	0
www.lafoga.org	20031101	0
www.lafoga.org	20031115	0
www.lafoga.org	20031201	0
www.mbuhari.com	20030501	1350147
www.mbuhari.com	20030703	1098316
www.mbuhari.com	20030716	1098316
www.mbuhari.com	20030801	1098316
www.mbuhari.com	20030815	1098316
www.mbuhari.com	20030901	1098316
www.mbuhari.com	20030915	1098316
www.mbuhari.com	20031001	1098976
www.mbuhari.com	20031015	1098976
www.mbuhari.com	20031101	1098976
www.mbuhari.com	20031115	1098976
www.mbuhari.com	20031201	1098976
www.otunbagbengadaniel.org	20030501	2199412
www.otunbagbengadaniel.org	20030703	2199412
www.otunbagbengadaniel.org	20030716	2199412
www.otunbagbengadaniel.org	20030801	2199412
www.otunbagbengadaniel.org	20030815	2199412
www.otunbagbengadaniel.org	20030901	2199412
www.otunbagbengadaniel.org	20030915	2199412
www.otunbagbengadaniel.org	20031001	2199412
www.otunbagbengadaniel.org	20031015	2199412
www.otunbagbengadaniel.org	20031101	2199412
www.otunbagbengadaniel.org	20031115	2199412
www.otunbagbengadaniel.org	20031201	2199412
www.rimionline.com	20030501	1179363
www.rimionline.com	20030703	1180165
www.rimionline.com	20030716	1179363
www.rimionline.com	20030801	1180165
www.rimionline.com	20030901	1179723

Curatorial Investigation

www.rimionline.com	20030915	1179723
www.rimionline.com	20031001	1179723
www.rimionline.com	20031015	1153970
www.rimionline.com	20031101	1153970
www.rimionline.com	20031115	1153970
www.rimionline.com	20031201	1153970
www.sarahjibril4president.org	20030501	863664
www.sarahjibril4president.org	20030703	863664
www.sarahjibril4president.org	20030716	863664
www.sarahjibril4president.org	20030801	863664
www.sarahjibril4president.org	20030815	863664
www.sarahjibril4president.org	20030901	863664
www.sarahjibril4president.org	20030915	863664
www.sarahjibril4president.org	20031001	863664
www.sarahjibril4president.org	20031015	863664
www.sarahjibril4president.org	20031101	863664
www.sarahjibril4president.org	20031115	863664
www.sarahjibril4president.org	20031201	863664
www.unongo.com	20030501	277572
www.unongo.com	20030703	277572
www.unongo.com	20030716	277572
www.unongo.com	20030801	0
www.unongo.com	20030815	277572
www.unongo.com	20030901	277572
www.unongo.com	20030915	277572
www.unongo.com	20031001	277572
www.unongo.com	20031015	277572
www.unongo.com	20031101	277572
www.unongo.com	20031115	277572
www.unongo.com	20031201	277572
www.anppusa.org	20030501	690079
www.anppusa.org	20030703	5186
www.anppusa.org	20030716	676280
www.anppusa.org	20030901	676280
www.anppusa.org	20030915	676280
www.anppusa.org	20031001	676280
www.anppusa.org	20031015	676280
www.anppusa.org	20031101	676280
www.anppusa.org	20031115	676280
www.anppusa.org	20031201	676280
www.buhari-okadigbo.com	20030501	596224
www.buhari-okadigbo.com	20030703	596224
www.buhari-okadigbo.com	20030716	596224
www.buhari-okadigbo.com	20030815	596224
www.buhari-okadigbo.com	20030901	596224
www.buhari-okadigbo.com	20030915	596224
www.buhari-okadigbo.com	20031001	596224
www.buhari-okadigbo.com	20031101	596224
www.buhari-okadigbo.com	20031115	596224
www.buhari-okadigbo.com	20031201	596224
www.eueomnigeria.org	20030501	3371195
www.eueomnigeria.org	20030703	4293231
www.eueomnigeria.org	20030716	4293231
www.eueomnigeria.org	20030801	4293231
www.eueomnigeria.org	20030815	4293231
www.eueomnigeria.org	20030901	4293231
www.eueomnigeria.org	20030915	4293231
www.eueomnigeria.org	20031001	4293231
www.eueomnigeria.org	20031015	4293230

Curatorial Investigation

www.eueomnigeria.org	20031101	4293230
www.eueomnigeria.org	20031115	4293231
www.eueomnigeria.org	20031201	4293231
www.unnigeriaelections.org	20030501	2810060
www.unnigeriaelections.org	20030703	3333069
www.unnigeriaelections.org	20030716	3333069
www.unnigeriaelections.org	20030801	3333069
www.unnigeriaelections.org	20030815	3333069
www.unnigeriaelections.org	20030901	3333069
www.unnigeriaelections.org	20030915	3882772
www.unnigeriaelections.org	20031001	4476275
www.unnigeriaelections.org	20031015	4476275
www.unnigeriaelections.org	20031101	4476275
www.unnigeriaelections.org	20031115	4476275
www.unnigeriaelections.org	20031201	4681841

HTTP server distribution

In the chart below, the column labeled www uses data for the World Wide Web as a whole, derived from the Netcraft survey in June 2002. That time corresponds to the beginning of data collection for both the arl and asia sets. Nigerian site data collection did not begin until April 2003. The chart compares http server distribution in the general Web with that in the arl, asia and Nigerian sites.

At start of surveying	www* (6/02)	arl # (6/02)	arl %	asia # (6/02)	asia %	Nigeria # (4/03)	Nigeria %
Apache	64.4	81	65.3	40	74.1	31	83.4
Microsoft	24.9	15	12.1	9	16.7	3	8.1
Netscape	1.7	20	16.1	1	1.9	0	0
Other	9.0	8	6.5	4	7.4	3	8.1
Total		124		54		37	

* According to Netcraft report for June 2002 (<http://www.netcraft.com/Survey/index-200206.html>)

The chart below compares http server distribution in the general Web with that in the arl, asia and Nigerian sites. All arl, asia and Nigeria data is from December 14, 2003. The comparative data from Netcraft is from approximately December 2003. By this time, some of the Asia and Nigeria are down, while others may no longer have their original content. Thus in a few cases, the server software recorded below may be running on a site not in our original cohort. It should also be noted that many, if not most of the Asia and Nigeria sites use commercial servers located out of country and for which they do not control the choice of server software.

As of 12/14/2003	www*	arl #	arl %	asia #	asia %	Nigeria #	Nigeria %
Apache	67.4	93	75.0	33	73.3	24	80.0
Microsoft	20.9	15	12.1	9	20	3	10.0
Netscape	3.3	9	7.3	0	0	0	0
Other	8.4	7	5.6	3	6.7	3	10.0
Total		124		45		30	

* According to Netcraft report for December 2003 (http://news.netcraft.com/archives/2003/12/02/december_2003_web_server_survey.html)

Below is a chart I created for the original JCDL paper. It shows the percentage distribution of Apache, Microsoft and Netscape http servers for the ARL sites as well as for all .edu sites, for three time periods.

HTTP server (by major category)	July 2002		January 2003		December 2003	
	% of 'edu' servers (77,869 surveyed)	% of ARL servers (124 surveyed)	% of 'edu' servers (approx. 78,455 surveyed)	% of ARL servers (124 surveyed)	% of 'edu' servers (approx. 99,814 surveyed)	% of ARL servers (124 surveyed)

Curatorial Investigation

			surveyed)		surveyed)	
Apache	51	67	52	68	54.5	75
Microsoft	31.5	12.5	33	14.5	33.5	12
Netscape	5.5	14.5	4.5	12	3.5	7
Other	12	6	10.5	5.5	8.5	6

* According to E-soft Inc. (http://www.securityspace.com/s_survey/data/200311/edu/index.html and related pages)

APPENDIX 35

Oracle Intermedia Full Text Search Implementation

Search Collections - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://dlibdev.nyu.edu/fasearch/crl/> Go Links >>

CENTER FOR RESEARCH LIBRARIES

CRL Political Web Archiving Project

**Search Web Sites
in the Political Web Archive**

Go to [Advanced Search](#)

Search for: In:

and/or/not

In:

Select a Region:

[Search hints](#) [About the database](#)

Sample searches: Agenda in all fields, Basque Country in place name, Aniagolu in person

For listings of all available websites by region [follow this link](#)

Start | The Digital Library | NYU - ... | Appendices - Microsoft W... | Search Collections - Mi... | Internet | 2:51 PM

APPENDIX 36

Sample Archive Records - MODS Descriptors

Sample Archive Record - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites

Address <http://dlibdev.nyu.edu:8083/xmldev/servlet/SaxonServlet?source=uraniagolu.xml&style=modspage.xsl> Go Links

Sample Archive Record

Title:	Loretta Aniagolu Web Site
Alternative Title:	Official Web site of Chief Loretta Aniagolu
Abstract:	Web site promoting the candidacy of Loretta Aniagolu, NCP candidate for governor of the Enugu State in the April 2003 Nigerian Election. Includes biography page, photo gallery, agenda, selected links
Creator:	Aniagolu, Loretta
Capture Date Range:	20030417-20030420
Subjects:	Elections NCP (National Conscience Party) Political Communications
Language:	eng
Genre:	Web site
Access Condition:	[Policy Statement Goes Here]
Active URL:	None
Archive:	CRL Political Web Archiving Project

Go to METS Viewer for Archived Versions of This Site:

[April 17, 2003](#)
 [April 19, 2003](#)
 [April 20, 2003](#)

WELCOME TO MY WEB SITE - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites

Address <http://dlibdev.nyu.edu:8083/xmldev/servlet/SaxonServlet?source=uraniagolu.xml&style=modspage.xsl> Go Links

Aniagolu turns on the HEAT Agenda, not Gender

YOU ARE INVITED
 Committee of Friends of
 Loretta Aniagolu invites
 you to a dinner to see
 her and hear signing of
 "ENIGUEN SPICES"

WHY? Holding her Legals
 Candidate Roll from 2002
 Street, U.S., Washington, DC
 (at the Capital)
DATE: Sat., March 29, 2003
TIME: 6:00 pm - 10:00 pm
RSVP: lor@aniagolu.com

MY MISSION
 To build an Enugu State with highly skilled, productive, and efficient manpower base, supported by adequate healthcare and food supply, secured in better lives and life styles.
 Change where superior service and commitment to excel bring about new level of performance.
 Change directed at harnessing resources, solving problems and creating opportunities for all.

AGENDA
 Health and Housing
 Education and Employment
 Agriculture and Arts
 Transport and Tourism

ALWAYS REMEMBER: In the eyes of God, all men and women are equal.

Vote for Change and enjoy a better tomorrow!

BOOK LAUNCHING
 A dinner/book signing in Washington, DC honoring
 Loretta Aniagolu's
 "Eniguen Spices"

Friday, September 19, 2003

Start | The Digital Library | NYU ... | Appendices - Microsoft W... | Sample Archive Recor... | Internet | 2:56 PM

Sample Archive Record - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Print Mail News

Address http://dlibdev.nyu.edu/webarchive/ibarretxemods.html

Sample Archive Record

Title: Juan Jose Ibarretxe

Alternative Title: Official website of Juan Jose Ibarretxe

Creator: Juan Jose Ibarretxe

Abstract: Ibarretxe.com is a website for supporting Basque Nationalism in the political arena and Juan J. Ibarretxe to continue being the Basque President of the Autonomous Community of the Basque Country in the European Union. The coalition of Basque Nationalist Party and Eusko Alkartasuna (Basque Solidarity) is supporting the team that have led the Basque Government in last decades, the team of the candidate Mr. Ibarretxe.

Capture Date Range: 09/07/2003 - 9/07/2003

Subjects: Political Communications--Western Europe
Basque Country
Internal Politics
Electoral Alliance
Election

Language: BAQ, SPA, ENG



Genre: Website

Access Condition: [Policy Statement Goes Here]

Active URL: http://www.ibarretxe.com

Archive: Political Communications Web Archive

Go to Archived Versions of This Site:

The screenshot shows a website with a header that reads "bizitza Bai vida errespetu" and "Ibarretxe da Baietzta". Below the header, there is a large banner with the text "Eskerrik Asko Euskadi, eta orain... lanera!". The page also features a map of the Basque Country and several tables of data, likely related to election results or campaign statistics.

Internet

Start | The Digital Library | Appendices - Micro... | Inbox - Microsoft O... | Sample Archive ... | Document7 - Micro... | 3:33 PM

s 34-37

SAMPLE Record in METS Viewer

aniagolu20030417.xml - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Print Mail News RSS

Address http://dlibdev.nyu.edu:8083/xmldev/servlet/frames/peoplesmandate.xml

Back to Cross-Collection Search
Title: Loretta A...
Date Captured: ...
Abstract: Web...
candidacy of Lo...
candidate for go...
State in the Apr...
Includes biogra...
agenda, selecte...

Back to Homepe...
Links
Link to Who Is...
Link to Agenda...
Link to Favorite...
Link to Photo G...
Link to Feedba...

Links
[Back to Homepage]
Bio and speeches
defence
Untitled Document
contactus
defence
Bio and speeches
flagbearers
guestbook
Pmp Homepage
Bio and speeches
Manifesto
marauders
membership form
News Release
Obasanjo1
Untitled Document
runningmate
SNC
404 Not Found
SouthSouth

Search for
[]
Search Clear

Search for
[]
Search Clear

Done


Start

Internet

Open in New Window Open in Full Frame

CRL Political Web Archiving Project METS Viewer

http://dlibdev.nyu.edu/webarchive/metstest/www.peoplesmandateparty.org/index.html

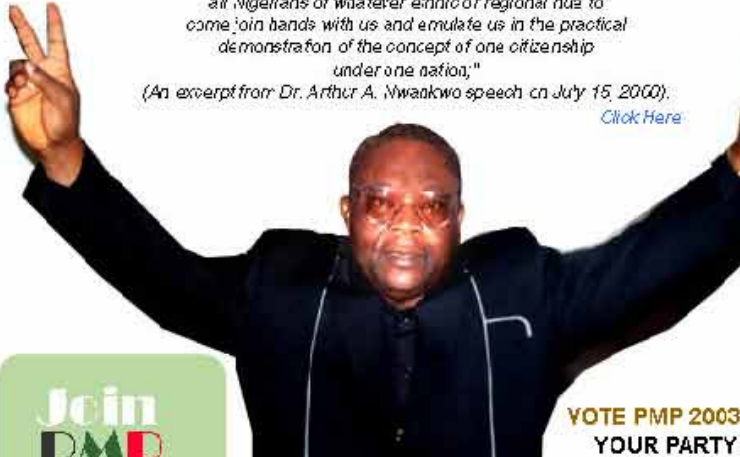
 **PMP** the choice of the
In 2003. Your Voice, Your Choice

Peoples Mandate Part

Dr. Arthur A. Nwankwo National Leader PMP
*"By the formation of this party we are presenting
for our people a platform on which to stan na and invite
all Nigerians of waaterer ethnic or regional hue to
come join hands with us and emulate us in the practical
demonstration of the concept of one citizenship
under one nation;"*
(An excerpt from Dr. Arthur A. Nwankwo speech on July 15, 2000).
[Click Here](#)

Be Thou Our Vision

- Home
- Constitution
- Manifesto
- Membership
- News Release
- Guest Book
- Photo Gallery
- News Headlines
- PMP-USA Who's Who
- Contact Us



Join PMP

**VOTE PMP 2003
YOUR PARTY**

Start | The Digital Library | Appendices - Micro... | Outlook Today - Mi... | http://dlibdev.n... | 3:29 PM

APPENDIX 38

Political Web Project - Investigation Wire Frame Draft per LTRM and Curatorial Team Meetings October 6, 2003

Purpose of the Archives

The archiving of Political Web materials will provide for the capture, preservation, and long-term availability for specific and limited educational and research uses, of important documents and messages disseminated by non-governmental political organizations and groups via the World Wide Web. These materials comprise a valuable source of information for historical studies, the social sciences, and public policy, but are by nature fugitive and susceptible to loss. Homogeneity of viewpoint in the current media [quote Carnegie report, Pew] parallels homogeneity of knowledge sources. Political Web is a thriving community of ideas and viewpoints, agendas whose artifacts are disappearing.

The primary uses of the archived materials are:

1. Scholarly research and teaching, in particular by historians and political scientists from all countries and regions
2. Study and informational use, by members of the international development, policy, diplomatic, and journalism communities, and lay individuals.

Archives Content

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

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

Related materials under the political communications rubric that might be addressed in subsequent investigations include:

- deep Web materials that are password-protected or otherwise designed to be resistant to robots and access
- electronic newspapers
- listserv digests
- RSS feeds; and
- databases.

Range of Archiving Activities

For purposes of the technical discussions the list below includes the working set of activities that would enable the assembly, preservation and accessibility of a persistent and inclusive archive of Web-based political communications. These activities are distinct from the governance activities, e.g., decision-making and transactions pertaining to scope of the archives, participants, dissemination, and disposal of archives. The list will be further refined, expanded, and revised as the investigation proceeds.

- Development and procurement of tools and techniques
- Determination and monitoring of archives scope
- Creation / specification / adoption of standards for archives content
- Prospecting for archives content
- 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
- Authorization of selectors
- Pointing / programming of Web crawler
- Undertaking Web crawl
- Capture of content and metadata
- Notification of content producer re archiving
- Indexing / metadata production
- Certification/documentation of integrity of content¹
- Archiving of master copy (“dark archives”)
- Archiving of fail-safe copy (backup archives)
- Storage / maintenance of bits
- Curating / structuring presentation of content
- Presenting content for viewing and searching (“light archives”)
- Managing / preserving resource
- Incorporation of other archives’ content
- Asset management - rights, funds, resources
- Management / re-aggregation of archives content
- Quality assurance of resource

¹ Objectively verifiable preservation of evidentiary information, relating to integrity of content, date, authorship, method of capture, “chain of custody,” changes made subsequently. Peculiar to political evidence, newspapers, government documents, and laws, statutes, far more than to e-journals, where updating of information is critical.

Appropriate Participants in the Archiving Activities - General Characteristics

Libraries, universities -- Because the investigation is driven by the needs of scholarly research and teaching, the primary participants in the archiving activities should be only those organizations and parties that are demonstrably responsive and accountable to the scholarly community. These include research libraries, universities, scholarly societies, scholarly publishers, and other organizations and parties that serve the humanities and social sciences disciplines. These parties' control of and involvement in the archiving activities should ensure that those activities are shaped and informed by the needs of the user communities. The mission-critical activities should also probably be undertaken under the control of the libraries.

The prospective archives are intended for study and informational purposes also by members of the international development, policy, diplomatic, and journalism communities and by lay individuals. Hence to ensure utility of the resource to this audience, institutes and other non-profit knowledge organizations, foundations, government bodies, and their agents might also take part in the archiving endeavor.

Scholars -- To ensure inclusion in the archives of the most important political Web content individual scholars, historians, and other specialists in the field will participate in selection and, to the some extent, in other archiving activities. The high level of sophistication and standardization required in the work of generating and maintaining such archives over the long term, and the importance of such archives as a resource for the larger scholarly community, however, recommends that some mechanism for certification of participating individuals be put in place.

Creators, Hosts - Because of the nature of their activities, for the political groups and organizations that use the Web as a medium, the archiving or permanence of their content is of secondary importance. For this reason the curatorial team concluded that the cooperation of creators in the archiving activities is not to be expected. The original producers or hosts of political Web content, hence, will be involved in archiving activities only to the extent that they permit (or prevent) inclusion of their content in the "light" or available archives. However, if deep Web material is eventually to be included in the archives cooperation from creators of some sites will be needed to permit physical retrieval of material, i.e., to circumvent technical protections like robot exclusions and password protection.

Roles -- The specific roles played by the respective participants in the archiving effort are determined by several factors. The intellectual property or copyright regime maintained by the archives effort, for instance, will affect the eligibility of various organizations and parties to play particular roles on the archiving. Taking a comprehensive approach to dark archiving Web materials, regardless of copyright, as the curatorial team has suggested, might limit the role that federal agencies like the Library of Congress or for-profit organizations, which are risk-averse with regard to copyright, might play in the archiving.

Moreover, copyright restrictions apply differently from one activity to the next. The highest risk associated with copyright infringement will inure to the access- and distribution-related activities. Conversely, risk will be lower for those engaged in curation and "dark" archiving activities.

Concern with restrictions like copyright will also inform the relationships developed among the participants. Since liability can transfer from one party to another in joint endeavors, the affiliation between the participating entities must be structured to insulate one party from liabilities incurred by another. Accordingly, the involvement of for-profit organizations in the archiving activity, for instance, should be limited to a vendor-client relationship strictly specified by contract stipulating a quid pro quo, rather than as a joint investment or partnership relationship.

The economic model developed (see below) will also affect the eligibility of parties to participate or benefit. For example, the federal support economic model might limit participation to entities within a specific country or geographic region.

General Characteristics of Resulting Resources and Distribution of Activities

The participants and activities identified above would contribute to development and maintenance of the archives as a resource for scholarship and research. Ideally, those activities would result in two kinds of resources:

1. A common or central resource, enabled by economies of scale and uniformity of practice and controlled collectively by the larger community of participating organizations, specifically libraries, universities, other non-profit knowledge organizations; and
2. Distributed or local resources developed with varying degrees of autonomy from the central resources but to some degree interoperable with the common resource through reliance upon tools, methodologies, standards, and other enabling mechanisms created and applied in the centralized archiving activities.

In order to achieve economies of scale some activities will ideally be centralized; others will be undertaken locally, to take advantage of the dispersed expertise and capabilities needed to support the archives. Ideally, the tools, methodologies, and management structures employed by the archives will support/enable both types of activities:

1. Centralized or core activities -- undertaken or controlled centrally to ensure integrity and persistence of the common resource
2. Distributed or local activities - These permit the archiving activities, particularly identification and selection of resources, to draw upon subject and technical expertise and capabilities where they reside; and to leverage investment already dedicated to development of comparable or useful resources, such as the creation and maintenance of subject portals; specialized archiving; comprehensive domain harvesting, etc. These activities are undertaken in conformance with standards and policies to augment the common resource but performed locally by designated parties individuals on a formal or informal basis.

Given the goal of the archives endeavor to ensure long-term availability of the resources developed critical management functions must be performed centrally where they would be most closely controlled by the community at large. Such critical functions might include rights procurement and management, identification of standards and policies, quality assurance, and so forth. Centralized programming would be advisable, for instance, as this is usually a costly activity. Other activities could be performed most economically or effectively by third parties, given the proper assurances and controls.

The ideal curatorial regime for targeting materials will probably combine manual and automated operations. One harvesting model discussed combines three elements: selection by curators; wholesale web crawling with smart, or "learning" robots; and reliance on larger, comprehensive Web archiving projects. Such a model might involve human and automated as well as centralized and distributed processes:

1. Curatorial / selection activities - identification of the characteristics (e.g., URL, domain, presence of keywords, etc.) of desired sites; specification of capture regime (periodicity, extent, etc.); negotiation of permissions (where appropriate). This might piggy-back on existing targeting and selection activities located where expertise and capability are already concentrated and supported to develop and maintain portals or subject/region-specific harvesting operations that gather political materials. Examples: region-based portal development at LC Portals, Africa portal at Stanford, LANIC Latin American portal/harvesting; also subject-based portals like Amnesty, Greenpeace, Georgetown U, etc. Also could build upon national-level activities to gather and archive nation-specific materials, e.g., as part of copyright repository functions undertaken by LC, National Library of South Africa.
2. Central/automated functions - utilities and services include: operates utilities, programs crawl, harvests, archives, aggregates; provides certification/ documentation of content authenticity;

Curatorial Investigation

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

3. Backup functions: The effort might rely on comprehensive harvesting efforts by organizations like the Internet Archive / Google / PANDORA / others to provide unrefined or raw resources for later culling or retrospective harvesting.

Prospective Economic Models

The LTRM team identified some possible models of economic sustainability to be considered by the investigation, with examples of projects or organizations that utilize such methods. The list will be refined and probably expanded during the course of the Mellon investigation, and the most appropriate model (or combination of models) specified.

- Fee-based / Subscription Aggregator - End-users or their agents (i.e., universities, libraries) support common resource through payments for site license or other form of controlled access to resource.
Examples: Project Muse, JSTOR, University press e-journals
- Federation / Consortium - Members of a community-of-interest formed around the common resource commit to long-term development and maintenance of the resource. Differs from subscription in that the community's interest extends beyond the specific resource, evinced by their role in governance.
Examples: RLG Cultural Materials, CRL, ARL/AAU Global Resources, OCLC Digital Cooperative, Scholarly Repositories (Berkeley Electronic Press with U California eScholarship Repository and NELLCO Legal Scholarship Repository, http://repositories.cdlib.org/escholarship/peer_review_list.html; D-Space?)
- Federal Support - Supported and controlled by a federal agency or organization or a federation of agencies/organizations, perhaps fulfilling a national depository function.
Examples: LC/National Digital Library, PANDORA, KulturarW3
- Philanthropic Support - Supported largely through grants and funding from government, university, and/or private philanthropy.
Examples: LANIC, H-Net, Internet Archive
- Hybrid - A combination of elements of the above models.
Examples: JSTOR, Ithaka, Internet Archive
- Broadcasting - "Some models are quite simple. A company produces a good or service and sells it to customers. If all goes well, the revenues from sales exceed the cost of operation and the company realizes a profit. Other models can be more intricately woven. Broadcasting is a good example. Radio, and later television, programming has been broadcast over the airwaves free to anyone with a receiver for much of the past century. The broadcaster is part of a complex network of distributors, content creators, advertisers (and their agencies), and listeners or viewers."
- Exchange - Emerging model of knowledge brokering, promoting on-line trading of information and opinion as represented by Opinion Exchange, DARPA (short-lived) exchange, others.

Some Additional Characteristics of the Prospective Archiving Endeavor:

- Adheres to guidelines for digital resource stability and interoperability set forth in the Open Archival Information System (OAIS) Reference Model.
- Adopts open source software to the extent possible and shares tools developed liberally. (This will make the resource more sustainable through wide adoption.) Services using the tools can be provided on a fee basis.
- Tools, methodologies are “platform-independent,” i.e., be adaptable to future technologies.
- Archives will be a growing resource rather than a finite “collection.”
- Access architecture should permit user authentication and varying degrees of control, to accommodate a subscription/membership economic model (for sustainability) as well as special access/privileges like in-country benefits for producer organizations/populations (Cf., Michigan State/ Project Muse African journals project as possible model).
- Capable of fluid but centrally controlled migration of content materials from “dark” archive to “light” archive, initiated by automated process and/or human intervention. (This allows archiving of materials whose presentation is temporarily prohibited by copyright, other circumstances. The original content producer/host will be able to opt out of inclusion in the light archives, and possibly the dark archives as well.)
- Scalability in accordance with varying levels of resources available.
- Preserves the integrity of content. The curatorial team advocates capture and/or documentation of as much of the formal and structural characteristics of the original presentation/content (“look and feel”) as practically and economically possible.
- Ability to partition among the various supporting activities to limit risk or exposure to parties engaged in certain activities.
- Capable of ingesting content from, or building on the selection activities of, other archives, repositories, comprehensive, like Internet Archives, Google, and region-specific selective like PANDORA, and Heidelberg (Chinaresource.org) effort on Chinese Web materials.
- Supports cross-regional/language as well a cross-disciplinary inquiry and research, including topical studies, such as human rights, immigration, etc.

To ensure availability of the archives content to the larger scholarly community, moreover, the archiving activities could not be wholly or even heavily reliant upon a single party, i.e., university or library, without certain guarantees.

+++++

APPENDIX 39

User Survey: Political Communications Web Archive

This survey instrument is part of a study on the preservation of political Web sites. It is designed to gather information about the use of political Web sites as primary source material in academic and policy research. The survey authors seek to determine how extensively Web sites are used for such research, and the needs, goals, and methodologies of the researchers.

For purposes of the survey, political Web sites are defined as sites produced and mounted on the Web by political parties, candidates, factions, and other non-governmental activist organizations and entities.

This survey is being conducted by the [Center for Research Libraries](#), in partnership with the [Latin American Network Information Center](#) at the University of Texas at Austin, Stanford University, Cornell University, New York University, and the Internet Archive. Funding is provided by the Andrew W. Mellon Foundation.

Your answers to this survey are confidential, and individual responses will not be shared with other parties unless required by the Texas Public Information Act. Aggregate data from this survey may be shared with third parties. Please consult the complete UT Austin [Web Privacy Policy](#) for additional details.

1. Affiliation:	
<input type="checkbox"/>	University Faculty
<input type="checkbox"/>	Graduate Student
<input type="checkbox"/>	Independent Researcher
<input type="checkbox"/>	Policy Institute
<input type="checkbox"/>	Government Agency
<input type="checkbox"/>	NGO
<input type="checkbox"/>	Other <input type="text"/>
2. Field(s) of study:	
<input type="checkbox"/>	History
<input type="checkbox"/>	Economics
<input type="checkbox"/>	Political Science
<input type="checkbox"/>	Sociology
<input type="checkbox"/>	International Relations
<input type="checkbox"/>	Public Policy
<input type="checkbox"/>	Literature and Language Studies
<input type="checkbox"/>	Religion
<input type="checkbox"/>	Environment
<input type="checkbox"/>	Other <input type="text"/>
3. Geographic region(s) of specialization:	

Curatorial Investigation

US/Canada

Latin America and Caribbean

Sub-Saharan Africa

Middle East and Northern Africa

Eastern Europe and Baltics

Western Europe

Central Asia

South Asia

Southeast Asia

East Asia

Australia/New Zealand

4. During the past two years, have you used materials from the Web in your research as citable sources of information?

Yes No

5. What percentage of your work time involves research?

6. What percentage of your research involves use of Web-based resources?

7. What percentage of your citations typically refer to Web-based resources?

8. What types of Web sites have you used:

Political party/candidate site

Protest/activist or social movement

NGO

Government

News service

Other

9. Have you accessed or monitored the same Web sites or materials repeatedly over time?

Yes No

10. If so, how frequently did you access those sites?

Daily

Weekly

Monthly

Other

11. Over what time period did you monitor those sites?

Curatorial Investigation

Days
 Weeks
 Months

12. What types of sites did you monitor?

Political party/candidate site
 Protest/activist or social movement
 NGO
 Government
 News service
 Other

13. Did you "archive" any of those sites?

Yes, I printed them out.
 Yes, I saved these sites to disk using the software.
 No, I didn't archive them at all.

14. What types of sites did you archive?

Political party/candidate site
 Protest/activist or social movement
 NGO
 Government
 News service
 Other

15. Are you familiar with the Internet Archive's [Wayback Machine](#)?

Yes No

16. If so, have you used the Wayback Machine to search for earlier versions of Web sites?

Yes No

17. If so, how often?

18. If so, how useful were the results for your research?

Very useful
 Somewhat useful
 Neutral
 Not particularly useful
 Not useful at all

Comments:

19. Which of the following non-Web-based materials have you used during the past five years as citable sources of information?

<input type="checkbox"/> Books	<input type="checkbox"/> Foreign broadcast reports
<input type="checkbox"/> Journals	<input type="checkbox"/> Monographs
<input type="checkbox"/> Newspapers	<input type="checkbox"/> Maps
<input type="checkbox"/> Government publications and documents	<input type="checkbox"/> Sound Recordings
<input type="checkbox"/> Pamphlets and broadsides	<input type="checkbox"/> Manuscripts
<input type="checkbox"/> Posters and ephemera	<input type="checkbox"/> Other <input type="text"/>

20. During the last two years has your use of any of these non-Web-based materials declined significantly?

<input type="checkbox"/> Books	<input type="checkbox"/> Foreign broadcast reports
<input type="checkbox"/> Journals	<input type="checkbox"/> Monographs
<input type="checkbox"/> Newspapers	<input type="checkbox"/> Maps
<input type="checkbox"/> Government publications and documents	<input type="checkbox"/> Sound Recordings
<input type="checkbox"/> Pamphlets and broadsides	<input type="checkbox"/> Manuscripts
<input type="checkbox"/> Posters and ephemera	<input type="checkbox"/> Other <input type="text"/>

21. If yes, is this due to your increased use of Web-based resources?

Yes No

22. Which domains do you find most useful to your research?

<input type="checkbox"/> .com
<input type="checkbox"/> .edu
<input type="checkbox"/> .gov
<input type="checkbox"/> .org
<input type="checkbox"/> .net
<input type="checkbox"/> .mil
<input type="checkbox"/> Other (i.e., .uk, .cn, .br, etc.) <input type="text"/>

23. Given the uncertain lifespan of Web sites, do you think the increased use of the Internet for political communications is problematic for future research?

Yes No

Comments:

24. How useful would an online archive of political communications Web sites be as a resource for research in your field?

- Very useful
- Somewhat useful
- Neutral
- Not particularly useful
- Not useful at all

25. For teaching your subject area?

- Very useful
- Somewhat useful
- Neutral
- Not particularly useful
- Not useful at all

26. What technical characteristics of the Web sites that you study or archive do you normally record?

- URL
- Date
- Server host IP number
- Software used with site content
- Authoring tools used
- Structural metadata
- URLs of linked sites
- URLs of sites that link to site studied
- Other

27. What content or information types from the Web sites that you study or archive do you normally record?

- Text
- Images
- Sound files
- Motion or Flash files
- Databases
- Other

Curatorial Investigation

28. Your name (optional):

29. Your email address (optional):

30. Please use the space below to include any additional comments:

This page was last updated on: 26 January 2004
[Center for Research Libraries](#) | [LANIC](#) | info@lanic.utexas.edu

APPENDIX 40

Political Communications Web Archive Survey, February 2004

General Notes

- There were 125 respondents to the survey.
- Because many questions were structured to allow participants to offer more than one response, most of the percentages add up to more than 100. Where participants could select more than one response, the raw number and percentage of participants who chose a single response are listed in the extended summary.
- In order to refer to the actual text of the questions, you may still view the survey at <http://lanic.utexas.edu/project/crl/survey/survey04.html>

Abbreviated Summary

- Affiliation: 46.4% faculty, 36.0% graduate student
- Most popular fields of study: History (39.2%), Political Science (35.2%), International relations (21.6%), Religion (19.2%)
- Most popular regions of study: Middle East and Northern Africa (65.6%), US/Canada (26.4%), Latin America/Caribbean (21.6%)
- 92.0% indicated that they had used Web resources as citable sources.
- Breakdown of Web site types used:
 - News service: 79.2%
 - Government: 74.4%
 - NGO: 66.4%
 - Protest/activist or social movement: 60.0%
 - Political party/candidate sites: 39.2%
 - Other: 36.8%
 - No answer: 4.8%
- 84.0% indicated that they had accessed or monitored the same Web site or materials repeatedly over time.
- Breakdown of Web site types monitored (several who declined to answer or answered “no” to the above question answered this one):
 - News service: 56.8%
 - Government: 45.6%
 - Protest/activist or social movement: 40.8%
 - NGO: 39.2%
 - Other: 26.2%
 - Political party/candidate sites: 24.8%
 - No answer: 8.8%
- 36.8% indicated they archived by printing, 28.0% did not archive, 16.0% archived by saving, another 13.6% archived by both printing and saving, and 5.6% did not answer this question
- Breakdown of types of sites archived:
 - News service: 46.4%
 - Government: 36.8%
 - No answer/none: 34.4%
 - NGO: 31.2%
 - Protest/activist or social movement: 27.2%
 - Other: 16.8%
 - Political party/candidate sites: 15.2%

Curatorial Investigation

- 84.8% indicated they were unfamiliar with the Wayback Machine, 12.6% said they were familiar with it, and 2.4% declined to answer
- 14 participants indicated they had used the Wayback Machine to search for earlier versions of Web sites. Of these, 8 indicated it had been “somewhat useful.”
- Books, journals, newspapers, and government publications and documents had all been used by more than 75% during the previous five years as citable sources.
- Declines in use of the following non-Web-based resources were indicated by more than 10% of participants: newspapers (24.8%), government publications and documents (16.0%), journals (16.0%), and books (12.0%).
- Of those who answered the question, 67.6% indicated that such declines were due to increased use of Web-based resources and 32.4% indicated that this was not the case.
- Breakdown of domains found most useful for research:
 - **.org: 80.0%**
 - .edu: 75.2%
 - .com: 56.8%
 - .gov: 53.6%
 - Other: 17.6%
 - .net: 28.8%
 - .mil: 5.6%
 - No answer: 5.6%
- 47.2% indicated they thought the increased use of the Internet for political communications was problematic for future research, 44.8% indicated that it was not problematic, and 8.0% did not answer. Of the 56 participants who indicated that it was not problematic, 4 noted that they had faith in archiving - 3 without mention of the medium on which sites should be archived, 1 singling out hard copies.
- 80.0% indicated that an online archive of political communications Web sites would constitute a useful resource for research in their fields, and 79.2% indicated that such an archive would be useful for teaching their subject areas.
- Regarding technical characteristics, 83.2% of participants indicated that they normally record the URL of Web sites that they study or archive, 60.8% record the date, and all other technical characteristics were mentioned by no more than 20% of participants.
- Regarding content or information types that are normally recorded, 94.4% selected “text,” 41.6% selected “images,” 22.4% selected “databases,” and each of the remaining options was selected by fewer than 10% of participants.
- There are many interesting individual comments that warrant perusal. Additionally, three participants requested that they be informed of findings.

Extended Summary

Affiliation:

- Faculty: 58 (46.4%)
- Graduate student: 45 (36.0%)
- Independent researcher: 5 (4%)
- Government: 3 (2.4%)
- Institute: 3 (2.4%)
- NGO: 3 (2.4%)
- Other: 8 (6.4%):
 - Community college faculty
 - Librarian: 2
 - Magazine researcher
 - Part-time university faculty
 - University-based researcher
 - UT PhD graduate student
 - Webmaster

2. Fields of study:

- History: 49 (39.2%)
- Economics: 11 (8.8%)
- Political Science: 44 (35.2%)
- Sociology: 21 (16.8%)
- International relations: 27 (21.6%)
- Public policy: 16 (12.8%)
- Literature: 15 (12%)
- Religion: 24 (19.2%)
- Environment: 7 (5.6%)
- Other: 35 (28%)
- *68 (54.4%) gave a single answer*

3. Geographic regions of study:

- US/Canada: 33 (26.4%)
- Latin America/Caribbean: 27 (21.6%)
- Sub-Saharan Africa: 6 (4.8%)
- Middle East and Northern Africa: 82 (65.6%)
- East Europe and Baltics: 3 (2.4%)
- Western Europe: 12 (9.6%)
- Central Asia: 6 (4.8%)
- South Asia: 5 (4.0%)
- Southeast Asia: 7 (5.6%)
- East Asia: 3 (2.4%)
- Australia/New Zealand: 1 (0.8%)
- *88 (70.4%) gave a single answer*

4. During the past two years, have you used materials from the Web in your research as citable sources of information?

- Yes: 115 (92.0% or 92.7%)
- No: 9 (7.2% or 7.3%)
- No answer: 1 (0.8% or n/a)

Curatorial Investigation

5. What percentage of your work time involves research?

- Ranged from 0% to 100%
- 1 answered "0%" (0.8%)
- 53 answered with a percentage between 0 and 50 or with a range that began below 50% and ended at 50% (42.4%)
- 34 answered "50%" (27.2%)
- 29 answered with a percentage above 50 and below 100 or with a range that began at 50% and ended below 100% (23.2%)
- 8 answered "100%" (6.4%)

6. What percentage of your research involves use of Web-based resources?

- Ranged from 0% to 100%
- 1 answered "0%" (0.8%)
- 97 answered with a percentage between 0 and 50 (77.6% or 78.2%)
- 9 answered "50%" (7.2% or 7.3%)
- 16 answered with a percentage above 50 - although the person who answered "59" may have mistyped "50" (12.8% or 12.9%)
- 1 answered "100%" (0.8%)
- 1 gave no answer (0.8% or n/a)

7. What percentage of your citations typically refers to Web-based resources?

- Ranged from 0 to 100%
- 4 answered "0%" (3.2% or 3.3%)
- 105 answered with a percentage between 0 and 50 (84.0% or 86.0%)
- 6 answered "50%" (4.8% or 4.9%)
- 6 answered with a percentage above 50 or with a range that began at 50% and ended below 100% (4.8% or 4.9%)
- 1 answered "100%" (0.8%)
- 3 gave no answer (2.4% or n/a)

8. What types of Web sites have you used?

- Political party/candidate sites: 49 (39.2% or 41.2%)
- Protest/activist or social movement: 75 (60.0% or 63.0%)
- NGO: 83 (66.4% or 69.7%)
- Government: 93 (74.4% or 98.2%)
- News service: 99 (79.2% or 83.2%)
- Other: 46 (36.8% or 38.7%)
 - No participants specified which other types of sites they have used.
- No answer: 6 (4.8% or n/a)
- *7 gave a single answer (5.6% or 5.9%)*

9. Have you accessed or monitored the same Web sites or materials repeatedly over time?

- Yes: 105 (84.0% or 86.8%)
- No: 16 (12.8% or 13.2%)
- No answer: 4 (3.2% or n/a)

Curatorial Investigation

10. If so, how frequently did you access those sites?

- Of those who answered “no” to question 9, 4 gave answers to this question: “Occasionally”; “Every now and then”; Daily; Weekly
- Of those who answered “yes” to question 9, all 105 answered this question:
 - Daily: 22
 - Daily and weekly: 4
 - Weekly: 41
 - Weekly and monthly: 4
 - Weekly, monthly, and other (“some annually”): 1
 - Monthly: 23
 - Other: 10 (other, “3-4 months,” “4 times a year, as needed,” “depends on the site and need,” “infrequently,” “irregularly & unscientifically,” “occasionally (verify cit.),” “once in a while, varies”)
- Of those who gave no answer to question 9, 3 gave answers to this question: Daily (2); “On and off depending upon need”
- 13 gave no answer

11. Over what time period did you monitor those sites?

- Of those who answered “no” to question 9, 4 (not the same 4 as above) gave answers to this question: Months (3); Weeks (1)
- Of the 105 who answered “yes” to question 9, 5 gave no answer to this question. The other answers were:
 - Days: 4
 - Days and a number: 0
 - Weeks: 6
 - Weeks and a number: 5 (2, 4, 6, 10, 10)
 - Months: 39
 - Months and a number, ranging from 1 to 120: 45
 - 1-12: 1 participant
 - 2: 4
 - 3: 2
 - 5: 2
 - 6: 4
 - 11: 1
 - 12: 9
 - 12-36: 1
 - 15: 1
 - 16: 1
 - 18: 1
 - 24: 8
 - 36: 2
 - 48: 2
 - 60: 2
 - 120: 3
 - Months years: 1
- Of those who gave no answer to question 9, 3 (the same 3) gave answers to this question: 2 days (2); Months
- 18 gave no answer

Curatorial Investigation

12. What types of sites did you monitor?

- Political party/candidate sites: 31 (24.8% or 27.2%)
- Protest/activist or social movement: 51 (40.8% or 44.7%)
- NGO: 49 (39.2% or 43.0%)
- Government: 57 (45.6% or 50.0%)
- News service: 71 (56.8% or 62.3%)
- Other: 33 (26.2% or 28.9%)
 - Of these, 26 specified which other types of sites. These included, for example, academic (3), all of the above (1), cultural sites (3), think tanks (2), religious sites (3), and “news media”/“online newspapers” (neither of whom chose “news service” above)
- No answer: 11 (8.8% or n/a)
- *24 gave a single answer (19.2% or 21.1%)*

13. Did you “archive” any of those sites?

- Yes, I printed them out: 46 (36.8% or 39.0%)
 - Of these, 1 printed them out and listed software under “saved,” below: Acrobat + Word
- Yes, I saved these sites to disk using the [box] software: 20 (16.0% or 16.9%)
 - Software used: not specified: 3; other answers included, but were not limited to, Adobe Acrobat, Netscape bookmarks, Internet Explorer, Microsoft, and Microsoft Word
- Yes, I printed and yes, I saved: 17 (13.6% or 14.4%)
 - Software used: not specified: 3; other answers included Acrobat, html, Internet Explorer, Netscape, and Word
- No, I didn’t archive them at all: 35 (28.0% or 29.7%)
- No answer: 7 (5.6% or n/a)
- *101 gave a single answer (80.8% or 87.8%)*

14. What types of sites did you archive?

- Political party/candidate sites: 19 (15.2% or 23.2%)
- Protest/activist or social movement: 34 (27.2% or 41.5%)
- NGO: 39 (31.2% or 47.6%)
- Government: 46 (36.8% or 56.1%)
- News service: 58 (46.4% or 70.7%)
- Other: 21 (16.8% or 25.6%)
 - Of these, 15 specified which other types of sites. These included “academic” or “specialist academic”(6) cultural sites (3), think tank (2), religious sites (2), and “news periodicals,” “online newspapers,” and “journal articles” (1 each, none of whom chose “news service” above)
- 43 gave no answer or answered with “-----” (34.4% or n/a)
 - Of these, 2 answered “yes” to question 13
 - Of those who answered “no” to question 13, all but one left this question blank. That participant answered “other,” unspecified.
- 18 gave a single answer (14.4% or 22.0%)

15. Are you familiar with the Internet Archive’s Wayback Machine?

- Yes: 16 (12.8% or 13.1%)
- No: 106 (84.8% or 86.9%)
- No answer: 3 (2.4% or n/a)

16. If so, have you used the Wayback Machine to search for earlier versions of Web sites?

- Yes: 14 (11.2% or 19.7%)
- No: 57 (45.6% or 80.3%)
- No answer: 54 (43.2% or n/a)
- All who answered “yes” to this question also answered “yes” to question 15
- Of those who answered “yes” to question 15, 2 answered “no” to this question and the remaining answered “yes”
- Of those who gave no answer to this question, 53 answered “no” to question 15 and 1 gave no answer to question 15
- The remaining 2 who gave no answer to question 15 answered “no” to this question

17. If so, how often?

- All 14 who answered “yes” to question 16 gave an answer to this question: “4-5 times/yr,” “half a dozen?,” “all the time,” “4,” “Every few weeks,” “Once. May do it more,” “6 or so times a year,” “dozen or so,” “every 6 months,” “Only a few times,” “half dozen times,” “half a dozen,” “10 x,” “rarely”
- All others gave no answers or indicated that the question was not applicable

18. If so, how useful were the results for your research?

- Again, all 14 answered, but several others answered as well. First, the 14 who said they had used the Wayback Machine to search for earlier versions:
 - Very useful: 3
 - Somewhat useful: 8
 - Neutral: 1
 - Not particularly useful: 1
 - Not useful at all: 1
- The remaining answers (other than “n/a”) *all came from people who claimed to be unfamiliar with the Wayback Machine:*
 - Very useful: 1
 - Somewhat useful: 2
 - Neutral: 2
 - Not particularly useful: 2
 - Not useful at all: 0
- With one exception, comments came from people who answered “yes” to question 16:
 - The Wayback archive is difficult to use, or at least it was when I used it. It was difficult to trace sites over time.
 - Problem with the existing archive is of course that it's far from comprehensive and even sites that interest me may have been dropped from being archived, without the reason being explained.
 - Hard to use. Requires more systematic approach than I have yet had time to develop. Anticipate that it will be v useful one day, as no alternative.
 - The Wayback Machine's archive is VERY spotty. Only selected pages are archived.
- The remaining comment, “I will read the information on this site and see how to use it,” came from someone who answered “no” to question 15 and left question 16 blank.

Curatorial Investigation

19. Which of the following non-Web-based materials have you used during the past five years as citable sources of information?

- Books: 118 (94.4 % or 96.7%)
- Journals: 116 (92.8 % or 95.1%)
- Newspapers: 98 (78.4 % or 80.3%)
- Government publications and documents: 88 (78.4 % or 72.1%)
- Pamphlets and broadsides: 52 (41.6 % or 42.6%)
- Posters and ephemera: 23 (18.4 % or 18.9%)
- Foreign broadcast reports: 45 (36 .0% or 36.9%)
- Monographs: 72 (57.6 % or 59.0%)
- Maps: 42 (57.6% or 34.4%)
- Sound recordings: 22 (17.6% or 18.0%)
- Manuscripts: 67 (53.6 % or 54.9%)
- Other: 13 (10.4 % or 10.7%)
 - Of these, 12 specified one or more kinds of “non-Web-based materials”:
 - Archival materials: 1
 - Dictionaries: 1
 - International agencies/UN or other international documents: 2
 - Interviews: 3
 - Oral histories/oral testimony: 2
 - Personal letters: 1
 - Videos: 1
 - Web articles and discussion lists: 1
 - Web sites: 1
- No answer: 3 (2.4% or n/a)
- *2 (1.6%) gave a single answer*
- It is clear that at least 2 participants answered with Web-based materials in mind; it is impossible to say how many of the other participants answered in a similar fashion.

20. During the last two years has your use of any of these non-Web-based materials declined significantly?

- Books: 15 (12.0% or 23.1% not considering those with no answer or 24.6% not considering others/no answers)
- Journals: 20 (16 .0% or 30.1% or 32.8%)
- Newspapers: 31 (24.8 % or 47.7 or 50.8%)
- Government publications and documents: 20 (16 .0% or 30.1% or 32.8%)
- Pamphlets and broadsides: 8 (6.4% or 12.3% or 13.1%)
- Posters and ephemera: 2 (1.6% or 3.1% or 3.3%)
- Foreign broadcast reports: 6 (4.8% or 9.2% or 9.8%)
- Monographs: 6 (4.8% or 9.2% or 9.8%)
- Maps: 1 (0.8% or 1.5% or 1.6%)
- Sound recordings: 3 (2.4% or 4.6% or 4.9%)
- Manuscripts: 5 (4.0% or 7.7% or 8.2%)
- Other: 4 (3.2% or 6.2% or n/a): “Increased my print-media usage” (1) and no/no decline (3)
- No answer: 60 (48.0% or n/a or n/a)
- *30 (24.0% or 46.2% or 49.2%) gave a single answer*

Curatorial Investigation

21. If yes, is this due to your increased use of Web-based resources?

- Yes: 50 (40.0% or 67.6%)
- No: 24 (19.2% or 32.4%)
- No answer: 51 (40.8% or n/a)
- All of those who gave no answer to question 20 either left this question blank (48) or answered “no” (12)
- All of those who answered “other” likewise either left this question blank (3) or answered “no” (1)
- There was no discernible pattern among the remaining 12 participant who answered “no” to question 21 and listed non-Web-based materials for question 20 or among the 50 who answered “yes” to this question, all of whom checked at least one box for question 20

22. Which domains do you find most useful to your research?

- .com: 71 (56.8% or 65.7%)
- .edu: 94 (75.2% or 87.0%)
- .gov: 67 (53.6% or 62.0%)
- .org: 100 (80.0% or 92.6%)
- .net: 36 (28.8% or 33.3%)
- .mil: 7 (5.6% or 6.5%)
- Other: 22 (17.6% or 20.4%), of whom 18 noted gave specific answers:
 - “All!”: 1
 - “etc.” (accompanied “.lb”): 1
 - “Middle Eastern sites”: 1
 - .au: 1
 - .br: 2
 - .co.uk: 2
 - .eg: 1
 - .fr: 1
 - .il: 4
 - .lb: 3
 - .le: 1
 - .ma: 2
 - .mk: 1
 - .tr: 1
 - .uk: 6
 - .un: 1
- No answer: 7 (5.6% or n/a)

Curatorial Investigation

23. Given the uncertain lifespan of Web sites, do you think the increased use of the Internet for political communications is problematic for future research?

- Yes: 59 (47.2% or 51.3%)
- No: 56 (44.8% or 48.7%)
- No answer: 10 (8.0% or n/a)
- 2 who did not answer this question said “maybe” or “I’m not sure. Not everything needs to live forever. Web sites have high percentages of junk and noise, with a few gems” in the comments section.
- 9 who answered “no” offered comments. Of these, 4 expressed the belief that the sites would be properly archived, virtually or on paper, and so posed no problem:
 - A suitable system for archiving news sources should be developed to resolve this surely.
 - As long as there is a hard copy version available in addition to the web site version. The web is for searching convenience, and is not "hard" source, but that is OK. For example, it makes a lot of sense to have journals on line in addition to their "hard" version. It is a good system because it saves a lot of time in searching, and you can be confident about what you download.
 - I trust we'll take the archiving need seriously.
 - Once it became clear that web-sites change or disappear completely, they can be archived.
- The remaining 5 did not express concern for archiving:
 - On the contrary, I think it is helpful.
 - political life is ephemeral and time-sensitive
 - Radio and other electronic media have posed similar problems for years - perhaps the internet has only increased our expectations that ephemera should all be recorded permanently!
 - The government doesn't archive their physical papers very well either and we, as citizens, rarely see the paper versions. I think the web is a good medium for communication for the government as well.
 - web site lifespan often has to do with how well it presents its material and keeps timely
- Of the 59 who answered “yes,” 29 offered comments, expressing concerns about authenticity, disappearance, and undocumented changes to sites

24. How useful would an online archive of political communications Web sites be as a resource for research in your field?

- Very useful: 62 (49.6% or 50.4%)
- Somewhat useful: 38 (30.4% or 30.1%)
- Neutral: 14 (11.2% or 11.4%)
- Not particularly useful: 9 (7.2% or 7.3%)
- Not useful at all: 0 (0.0%)
- No answer: 2 (1.6% or n/a)

25. For teaching your subject area?

- Very useful: 45 (36.0% or 38.1%)
- Somewhat useful: 54 (43.2% or 45.8%)
- Neutral: 15 (12.0% or 12.7%)
- Not particularly useful: 4 (3.2% or 3.4%)
- Not useful at all: 0 (0%)
- No answer: 7 (5.6% or n/a)

26. What technical characteristics of the Web sites that you study or archive do you normally record?

- URL 104 (83.2% or 92.9%)
- Date 76 (60.8% or 67.9%)
- Server host IP number 5 (4.0% or 4.5%)
- Software used with site content 2 (1.6% or 1.8%)
- Authoring tools used 5 (4.0% or 4.5%)

Curatorial Investigation

- Structural metadata 4 (3.2% or 3.6%)
- URLs of linked sites 25 (20.0% or 22.3%)
- URLs of sites that link to sited studied 18 (14.4% or 16.1%)
- Other 5 (4.0% or 4.5%):
 - author, organization, language
 - author, sources
 - DOWnload accesible material
 - Overall Format, Content
 - Sometimes some of the rest of the above, too (etc) [answered URL, date, metadata]
- No answer 13 (11.6% or n/a)

27. What content or information types from the Web sites that you study or archive do you normally record?

- Text: 118 (94.4% or 100%)
- Images: 52 (41.6% or 44.1%)
- Sound files: 10 (8.0% or 8.5%)
- Motion or Flash files: 5 (4.0% or 4.2%)
- Databases: 28 (22.4% or 23.7%)
- Other: 2 (1.6% or 1.7%):
 - All relevant
 - maps
- No answer: 7 (5.6% or n/a)

28. Your name (optional): 42 provided a name

29. Your email address (optional): The same 42 provided e-mail addresses, along with another 4

30. Comments:

- 22 gave comments, 1 of which was simply “I hope this helps in your endeavour,” 1 of which was a thank-you, and another 3 of which were requests that we inform the participant of findings.
- The remaining 17 comments are worth reading and can be found in the spreadsheet

31. Remote hosts: data were captured for all but 7 participants.

ⁱ See the official METS web site hosted by the Library of Congress, <http://www.loc.gov/standards/mets/>.

ⁱⁱ These challenges have been outlined in Peter Lyman’s article in the NDIP Plan: http://www.digitalpreservation.gov/ndiipp/repor/repor_plan.html as well as in any number of preliminary reports for various web archiving endeavors. See, for instance the netarkivet.dk final report <http://www.netarkivet.dk/rap/webark-final-rapport-2003.pdf> or the DAVID report at <http://www.dma.be/david/teksten/Report5.pdf>

ⁱⁱⁱ <http://nwa.nb.no/aboutNwaT.php>

^{iv} <http://nwa.nb.no/nwadocformatxmlschema.php>

^v <http://www.netarkivet.dk/rap/webark-final-rapport-2003.pdf>

^{vi} <http://www.clib.org/pubs/reports/pub106/web.html>

^{vii} <http://bibnum.bnf.fr/ecdl/2002/lc/lc.html>

^{viii} The Political Communications Web Archiving Project is a Mellon-funded research and planning initiative under the coordination of the Center for Research Libraries (CRL). This project will lay the methodological groundwork for the cooperative preservation of important documents and messages disseminated via the World Wide Web by non-governmental political groups and parties. The joint planning effort focuses on four world regions, each under the responsibility of one 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).

Curatorial Investigation

^x <http://www.library.cornell.edu/iris/research/WebPolCom.pdf>

^x Brewster Kahle et al, WWW Archive Format Specification
<http://pages.alexandria.com/company/arcformat.html>