Political Communications Web Archiving

A Proposal to the Andrew W. Mellon Foundation
from the
Center for Research Libraries

July 26, 2002

Bernard F. Reilly, Jr., Project Director

Political Communications Web Archiving

Project Summary

The Center for Research Libraries seeks funds for an investigation and planning effort to develop effective methodologies for the systematic, sustainable preservation of Web-based political communications. The effort will focus on Web materials produced by political groups and NGOs in Latin America, Sub-Saharan Africa, Southeast Asia, and Western Europe. These materials include reports, manifestos, constitutions and declarations, official statements, and other documents issued via the Web by individual political activists, political parties, and popular front and radical organizations. Such communications are vital primary source materials for history and area studies, but tend to be produced erratically and disappear quickly. The proposed undertaking, a combined effort of the Center for Research Libraries, four U.S. universities, and the Internet Archive, will lay the groundwork for cooperative preservation of these materials.

The project will unfold in three "segments," which will explore key aspects of political Web communications archiving:

1. *Curatorship* -- The optimal curatorial regimes, practices, and tools for the ongoing identification, targeting and capture of the various types of Web political communications to be archived.

2. *Technology* -- The general technical requirements, specifications, and tools best suited to the capture and archiving of political communications.

3. *Long-term resource management* – the organizational and economic framework necessary to support the cooperative archiving, management, and preservation of Web political materials on an ongoing basis.

The planning effort will draw upon the expertise of technology and subject specialists and scholars at New York University, Cornell University, Stanford University, and the University of Texas at Austin. Each university has outstanding collections and curriculum strength in the history and languages of a specific targeted geographic region, and has already begun to explore the capture and archiving of political Web materials from that region. (Statements outlining the political Web digital preservation investigations underway at each of the participating universities are appended.)

The proposed cooperative effort will build upon the investigations currently underway at these universities, and at the Library of Congress and the Internet Archive, and will draw conclusions and identify methodologies that can be generally applied by the larger research library community and across regions. In this way, the benefits of these investigations will accrue to the larger scholarly research community and to the national preservation effort.

Eastern European, African, and Latin American areas studies specialists from the Library of Congress will advise on issues of selection and curatorial methodology, and will participate in the final evaluation of capture and curatorial regimens. The Library's National Digital Library technology staff will also participate in the archiving and delivery discussions. (The Library will subsidize the expenses and in-kind costs of staff participation in the project.)

Also participating will be Brewster Kahle and Michele Kimpton of the Internet Archive organization. The Internet Archive will provide the project a rich test bed of archived Web political communications in the periodic "snapshots" of the Web that it has been gathering since 1996. The organization will also make its robust Web "harvesting" capabilities and its large capacity for digital storage and archiving available for study by the project team as one of several possible technology platforms for archiving and delivery.

The Center for Research Libraries will provide an organizational umbrella for the project, building upon its successful record of cooperative development and preservation of scholarly resources from the major regions of the developing world. Members of the Center's Area Studies Council (ASC) will advise in evaluating the study's conclusions, providing a perspective augmenting those of the participating universities. The ASC will also be the primary mechanism of communication with the broader library and scholarly communities involved in international studies regarding the project's importance, goals, and results. As part of this effort the Center will explore and define ways in which it can provide a stronger framework for the cooperative gathering and preservation of international materials in the digital realm.

In general, the effort seeks to balance the autonomous efforts of the participating universities and organizations – shaped and energized by the "communities of interest" coalescing around region- and culture-based studies – with the needs of the larger scholarly research community.

Background

Over the last decade the World Wide Web has become a vital medium of political communication in many regions and countries. Such communications include manifestoes, constitutions and declarations, agreements, and other Web-mounted public documents; Web sites and electronic newsletters of particular political candidates or groups; and issue-specific email messages, discussion threads, and electronic forums. Examples of this kind of material include the Web page of the Liberation Army of the Free Papua Movement in Indonesia (**http://www.eco-action.org/opm/**); the *Valley Of 1000 Hills Declaration* of the Conference on

Community Rights held in Natal, South Africa (http://valleyof1000hills.cua.edu/); and the *Forcas de Libertacao de Estado de Cabinda's* declaration of the state of Cabinda's independence from Angola (http://www.cabinda.org/anglais.htm).

Such communications are the digital-era counterparts of the posters, pamphlets, and other forms of street literature that have long served as indispensable sources of information on political trends, ideologies, and activities. Unfortunately, the content of Web-based political communications is disappearing without being properly archived.

Despite their importance for scholarly research in the humanities and social sciences, political communications have received little attention in national-level efforts to devise a preservation strategy for digital materials. Those efforts have tended to focus on materials like electronic journals, moving image, audio, and commercial broadcast media, which are largely generated by commercial producers and organizations and where preservation strategies can involve the cooperation of the producers. Political communications, on the other hand, tend to be more ephemeral precisely because of the inherent difficulty of engaging the producers, who are usually political activists with no cognizance of, or interest in, the long-term archiving of their materials.

Nonetheless, various third-party efforts have been made to provide scholarly access to these kinds of resources. The most commonly adopted access method involves gathering and presenting through a single Web site, i.e. a "portal," links to sites pertaining to particular regions or political issues. Examples of this approach include the Library of Congress's *Portals to the World* (http://www.loc.gov/rr/international/portals.html), the University of Texas *Electoral Observatory* (http://lanic.utexas.edu/info/newsroom/elections/) and Stanford University's *Africa South of the Sahara* (http://www-sul.stanford.edu/depts/ssrg/africa/guide.html). Such portals do not, however, truly archive Web materials, but rather merely provide real-time access to resources that are currently available but which are likely to change and even disappear over time.

Another approach is large-scale harvesting of Web pages, pioneered by the Internet Archive, the National Library of Australia in its *Pandora* project, and the National Library of Sweden through the *Kulturarw3* project. These initiatives periodically sweep the Web, gathering broad categories of Web content into massive archives and make those archives available for use to researchers through customized viewing devices. The Internet Archive takes periodic "snapshots" of the entire Web. The Australian and Swedish projects continually harvest and archive all Web resources produced in their respective countries. These efforts are only broadly selective, and offer neither full curatorial evaluation of content for inclusion nor mediation in the presentation of that content.

The long-term and systematic preservation of Web-based political communications presents a special set of challenges and opportunities. The research value of these kinds of materials hinges upon the preservation or recording of certain "evidentiary" traits or features associated with the original presentation of the material on the Web. Such traits include date and time of initial presentation, authorship and/or source of transmission, the format of their original presentation, and others. The loss or absence of these attributes will compromise the documentary value of the materials.

On the other hand, political Web materials are exempt from certain difficulties encountered in archiving of other kinds of Web materials. Political communications, for instance, tend to be technologically relatively simple and self-contained, consisting of static HTML, PDF, and image source files. More complex digital objects like e-journals and e-books often incorporate databases, software, and other dynamic content, or generate presentation files "on the fly," thus complicating the task of capture and archiving.

Political Web communications, moreover, are less often restricted by proprietary considerations than other kinds of Web resources. The entities that produce such materials are often loosely affiliated groups not formally incorporated, and tend to desire the widest possible distribution of their materials. Hence, the use of such materials in an archival or educational context is unlikely to provoke complaint. (This understanding will, however, be tested in the curatorial investigation segment of this project.)

While each region and each type of material presents its own unique challenges, there are certain common obstacles to preserving Web-based communications in general. These include the scarcity of funding available to libraries for preserving primary source materials for non-U.S. studies, the lack of technical standards for such activity, and the dissimilarity of Web archiving to traditional collection development methodologies. A number of centers at U.S universities have begun to face these challenges individually. These include Cornell University (for Southeast Asia), New York University (for radical groups from Western Europe), Stanford University (Sub-Saharan Africa), and the University of Texas at Austin (Latin America). Given the scale of the challenges, however, it is clear that an effective solution must be a cooperative one. The Web political communications project will take advantage of the individual efforts to generate benefits for the larger community.

Project Activities

The proposed planning project will consist of three distinct but interlocking work segments or investigations, which will examine technical, curatorial, and long-term resource management aspects of Web archiving. Each segment will be centered at one of the participating institutions and will have its own team of participants. (See *Project Organization and Timeframe*, below.) The first segment will run for the entire project period, a timeframe of fifteen months. The second and third segments will be concurrent investigations, running for months two through fourteen. (See *Timeline*, appendix 5)

1) Long-Term Resource Management. The first and primary investigation will determine the organizational and economic framework necessary to support the archiving of Web political materials on an ongoing basis. The activities involved in such archiving include selection and targeting of important Web sites, capture or harvesting, storage and maintenance of the digital files, presentation and dissemination of archived content, and provision of fail-safe repository for same. Libraries do not now incorporate these activities in their collection development and preservation programs.

Some of these activities, such as selection and targeting, might be best managed locally, where the requisite area studies and bibliographic expertise reside. Others might be most efficiently performed centrally, by an organization or entity possessing robust specialized capabilities such as server support and file management. Investigators will determine where the responsibilities for such activities are best situated and the ideal configuration of relationships and partnerships needed to ensure that those responsibilities are fulfilled.

At the outset of the project investigators will provide an initial "wire-frame" set of expectations regarding the long-term archiving activities and the resources developed through those activities, to guide and inform the work of the curatorial and technical investigations. (See below.) These expectations pertain, for instance, to the intended end users and beneficiaries of the archived resources and the prospective allocation of responsibilities for capture, archiving, and dissemination of the resources. They will also encompass the acceptable funding models (such as subscription access or consortium-supported funding) that might support those activities.

During the project year investigators will analyze and assess the capabilities of the participating organizations, including the Center for Research Libraries area studies programs, regional studies centers at various universities and the Library of Congress to determine what roles that each might play in the cooperative archiving of Web political communications. Investigators will also examine existing not-for-profit digital archiving and distribution mechanisms (such as the Research Libraries Group's Cultural Materials Initiative, Johns Hopkins University's Project Muse, JSTOR, the Library of Congress's National Digital Library, University of Chicago Press, and the Internet Archive) as possible frameworks for dissemination of political communications Web resources.

Investigators will analyze the capture and archiving models recommended by the technical and curatorial investigations. On the basis of these findings investigators will identify the specific set of activities and the resources required to support long-term archiving and accessibility of political web materials on a cooperative basis.

Finally, investigators will then describe the cooperative framework and the concrete steps necessary to implement the capture, archiving, and resource management models in a way that will ensure ongoing support of the effort and equitable access for the larger community.

2) Curatorship. Since the Web offers a relatively unrestricted channel of political communication, the explosion in such communications makes identification and selection of the most important materials a daunting challenge. Using Latin America and Southeast Asia as separate test beds of Web political communications, investigators will evaluate and identify the appropriate curatorial regimes and necessary tools for selection of materials to be archived (e.g., frequency of capture, extent of pages captured, formats for long-term retention, metadata requirements, and so forth) and the costs, benefits and risk factors associated with those regimes. The selection and archiving methodologies will be tailored to the various types of political communication behaviors, inherent characteristics of the materials produced by political groups, and the requirements of historical and language studies scholarship for which these communications serve as "primary source" research materials.

In particular this investigation will build upon the efforts of area specialists at the University of Texas and Cornell University. Political Web sites generated by Latin American and Southeast Asian political groups and identified by Texas and Cornell specialists will be gathered and will serve as a diverse test bed for the investigation. Specialists at Texas, for instance, have learned much about the ideal frequency for capture of political sites, based on events such as elections, civil disturbances, speeches, and other occurrences that trigger changes in such sites. The proposed investigation will determine which conclusions and practices are generalize-able for the capture of materials across all regions. The Internet Archives will provide web-crawling, harvesting and search support for this effort.

The curatorial investigation will also draw upon the experience gained by Stanford University researchers in developing and maintaining the *Africa South of the Sahara* portal. Stanford's African Studies curator and resident scholars have considerable first-hand knowledge of the rigors and the issues connected with identifying and selecting political Web resources on an ongoing basis.

Here the investigation will also address proprietary considerations in the archiving of political materials. Stanford's curator has worked closely with African NGOs and libraries on questions of indigenous rights and the use of intellectual property. This is a realm in which copyright law provides little guidance, and the non-legal rights and concerns of indigenous parties must be considered. (This concern was also expressed in connection with NYU's exploration of European radical groups Web productions.) While the capture and archiving methodologies developed under this project will remain within the bounds of fair use, potential benefits to the originating regions or parties will be explored in formulating the optimum economic model.

3) Technology. The technical investigation will build upon inquiries underway at New York University and Cornell University. (See Appendices 2 and 3.) The primary focus of this segment will be to analyze and evaluate existing approaches that may be adapted to the capture and archiving of Web-based political materials for scholarly purposes. For example, the Internet Archive has proposed an archival format for Web sites and has adopted a comprehensive "snapshot" approach in its production of the *Election 2000*, *September 11* and *Election 2002* web archives. The National Library of Australia's *Pandora* project and the Royal Library of Sweden's *Kulturarw3* project use guided "robotic Web crawlers" to locate Web resources that match pre-determined characteristics, selectively download and convert those Web sites to specified file formats, and archive the resultant files.

Cornell has been evaluating the creation and adaptation of Web-based tools by the Internet Archive and other organizations and applying those tools to a study of Web resources that is the basis for their preservation risk management approach. The Southeast Asia sites are the basis of a case study on materials at risk that explores the custodial end of the preservation risk management spectrum. During the investigation Cornell University researchers will work closely with the Internet Archive to examine their proposed archive format and features of their capture, archiving, and presentation methodologies. New York University has begun a study of the techniques employed in the *Pandora* model for application to political materials, and will further investigate the similar effort at the Swedish Royal Library.

Investigators will identify the features of each of these and other approaches that best address the peculiar characteristics of political Web materials, and will identify the technical requirements, characteristics, costs, benefits, and risk factors associated with each. Investigators will recommend the best practices, methods, techniques, and tools needed to manage these resources over time and incorporate emerging technologies. These elements will comprise a framework adaptable by different universities for a modular and flexible approach to archiving political Web materials.

Such a framework might leverage established capture and archiving efforts, such as those of the Internet Archive or the National Library of Australia, or be a hybrid that incorporates aspects of both approaches, as well as other tools and techniques that are available for various aspects of Web management. At minimum, such a framework will adhere to the guidelines for digital resource stability and interoperability set forth in the *Open Archival Information System (OAIS) Reference Model* (Washington, DC: Consultative Committee for Space Data Systems, July 2001).

Project Organization and Timeframe

Throughout the planning effort, the Center for Research Libraries will coordinate activities and facilitate intercommunication among the participants. Each of the three interlocking investigations will be undertaken by a team consisting of representatives of the participating universities. CRL will hold a conference of leaders and principals of the project teams at the beginning of the project period; host a summary meeting for each team at the conclusion of their respective investigation; and host a whole-project plenary meeting at the conclusion of the project. The last will also include members of the Center for Research Libraries Area Studies Council.

Investigation 1: Long-Term Resource Management -- Bernard Reilly and James Simon, of the Center for Research Libraries, will lead this investigation. The team will also include leaders of the curatorship and technology teams, as well as Nancy McGovern (Cornell), Karen Fung (Stanford), and Carolyn Palaima (Texas). With day-to-day administrative and technical support provided by a project coordinator and CRL systems analyst respectively, Reilly and Simon will work on an ongoing basis with the leaders of the curatorship and technology teams to coordinate activities, communicate useful information, organize and convene meetings, evaluate findings, and report and publicize results.

Simon will also serve as liaison between the project and the Center's Area Studies Council, and will seek their advice on issues as appropriate. Reilly will be liaison with potential partner organizations for dissemination of the political archives, such as the Research Libraries Group, Project Muse, and others. Reilly will also explore the potential for collaboration and convergence with the government documents archiving project proposed by the University of California’s California Digital Library.

Investigation 2: Curatorship -- Latin American area studies specialist Carolyn Palaima at the University of Texas will lead this investigation, working closely with Southeast Asian and African curators Allen Riedy (Cornell) and Karen Fung (Stanford). Kent Norsworthy at Texas will provide digital resource management support for the effort, and will work with the metadata analyst at Cornell to determine metadata requirements and to model optimum administrative and technical metadata for captured sites. Graduate researchers at Stanford and Texas and Cornell content specialist Allen Riedy will compile sites and other Web resources and compare these with past versions of the same sites harvested by the Internet Archive. The programmer and systems engineer at the Internet Archives will work with the technology team to provide customized Web crawling services and to capture Web political content on a repetitive basis, to provide a critical mass of test of content for curatorial team analysis.

Carolyn Palaima (Texas) and Karen Fung (Stanford) will act as liaisons to political groups with whom they have longstanding contacts to ascertain acceptable terms for archiving materials when necessary. Peter Lor, National Librarian of South Africa, will advise the project on intellectual property issues from the perspective of indigenous groups.

The team leader and specialists will work together to develop and evaluate the curatorial methodologies and will confer with Latin American and African area studies specialists and legal counsel at the Library of Congress, drawing upon the expertise and experience gained by the latter in developing and maintaining its portal of international digital materials. The curator of the Tamiment Library will also participate in the curatorial investigation, evaluating recommended methodologies and tools with respect to their adequacy for Western European Web materials.

Team 3 – Technology— NYU's Digital Library Analyst will lead the technology team, working closely with Cornell University's Digital Projects Librarian and Nancy McGovern, and Michele Klimpton of the Internet Archive. The NYU Digital Library Analyst and programmer analysts at Cornell will analyze and assess the capture and archiving tools and techniques employed by *Pandora*, *Kulturarw3*, Internet Archives, and other entities and report on their properties and characteristics, and recommend an OAIS-compliant preservation metadata schema for Web sites.

The systems administrator engineer and programmer at the Internet Archives will support the work of the team by providing customized Web crawling services targeting materials per characteristics defined by technology and curatorial teams and capturing Web political content on a repetitive basis, thus ensuring a critical mass of test of content for analysis. This content will be placed on servers at Cornell and NYU.
The Center for Research Libraries systems analyst will advise team leaders with respect to CRL capabilities and requirements for providing fail-safe repository services and other possible support to the long-term archiving effort.

CRL will then work with participants and advisors to report on the project and its conclusions. The elements of the report will include:

- General specification of the most appropriate technology architecture(s) and techniques for gathering and preserving political communications, with associated costs, benefits, characteristics, and risk factors.

- An acquisitions and growth plan, reconciling Web archiving and curatorial methodologies with traditional collection development activities and the regimens and periodicities appropriate to the capture of various kinds of communications.

- Specification of the optimal design and requirements of the organizational model(s) necessary to support ongoing digital collection development and the long-term availability of the archived resources. Such specifications will indicate the costs and requirements of sustaining the underlying activities and infrastructure, and the roles and functions of organizations that might fulfill centralized functions, such as fail-safe repositories, underwriters, and distributors.

The report will detail the points of consensus and convergence of project investigators, recommend implementation steps, and suggest ways in which the models and best practices developed might be applied to additional categories of political communications, such as message strings, online forums, and so forth. The report will also explore points of convergence with the government documents archiving project at the University of California's California Digital Library. Recommendations will also cover the Center for Research Libraries' long-term role vis a vis area studies in the digital realm.