

**Web-based Government Information Project:
A Mellon Funded Initiative of the California Digital Library**

**Environmental Scan: Preliminary Survey Results
(ver. 3.2)**

Patricia Cruse
California Digital Library

Chuck Eckman
Stanford University Libraries

March 12, 2003

Contents

Introduction & Report	2
User's Functional Needs and Tools: Appendix A	8
Interview Question Set A: Appendix B	10
Project Level Interview Question Set: Appendix C	11
Individuals Consulted: Appendix D	12
Related Projects and Programs: Appendix E	13
Web-based Government Information Project--FAQ: Appendix F	19
Crawl Report and .gov Demographics: Appendix G:	21

Introduction

The California Digital Library, with support from the Andrew W. Mellon Foundation, is conducting a cost-benefit review of technologies and approaches appropriate for the capture, curation, and persistent management of web-based documents of US state and federal governments. This project is based on three fundamental assumptions. The first is that government information – whether created and distributed in analog or digital form – is a critical resource for academic scholarship, scientific research and citizen participation. The second is that web-based government information is increasingly content that is at-risk due to a variety of well-documented pressures.¹ The third is that memory organizations (libraries, archives and museums) must retain their historic roles as acquirers, organizers and preservers of content in the new networked digital environment.

The project's ultimate goal is to outline the requirements for stable and sustainable digital collection building for this genre of information. In order to achieve this, the project is engaged in several inter-related sets of activities including:

- (1) review the scope of the domain of web-based government information, focusing on the development of meaningful data such as size, diversity of formats and rate of growth; [Appendix G]
- (2) identify promising technologies and projects related to the capture, management, and preservation of this material; survey the managers of these technologies and projects to build an information base for the project; and
- (3) capture a subset of web documents from government sites to serve as a test bed for analyzing the technical requirements for building collections of web-based government documents.

In conjunction with the goal to build an information base derived from existing practices in capturing and preserving web-based government information, project staff initiated two sets of interviews. The first set of interviews is being conducted with leaders in the digital preservation and government information fields to help shape the nature of our inquiry; the questions addressed to this group are attached as appendix B. The second wave of interviews includes individuals that are directly associated with projects addressing some aspect of the problem; these questions are listed in appendix C.

The subjects of these interviews are listed in appendix D. A project list—designed to track existing projects and identify potential interview candidates—is included as appendix E. A set of Frequently Asked Questions (FAQ) was developed as a response to questions raised to project staff by interview subjects about the project, its terminology and assumptions. This FAQ is attached as appendix F.

Although project staff is still conducting these surveys, we believe that the information gained from early respondents point to a clear set of common themes and experiences. This interim report is intended to share these results with a broader community. We have broken this report into four sections: (1) general themes; (2) capture; (2) archive management; and (3) preservation.

¹ Wiggins, Richard. "Nine Modes of Digital Death" in Digital preservation: paradox and promise. *Library Journal*, April 15, 2001

General Themes

The overall themes outlined below do not constitute an exhaustive list of the observations we obtained both from leaders in the field and project managers during the course of our interviews. However, these themes provide a useful frame of reference for the project-based interviews.

Challenges to automated or "bulk" collecting

Interviewees identified drawbacks of bulk-collecting, crawler-based approaches: (1) databases, password protected, and other deep web content is mostly missed by web crawlers; (2) government agencies often retain a variety of digital formats that may be not present on the web; (3) capture may involve too much throw away items if not carefully targeted, and targeting is not well-supported among existing technologies.

Challenges to selective or "by-hand" collecting

Several respondents clearly articulated two distinct disadvantages involved in selective or by-hand endeavors of capturing web-based materials: (1) the store of web-based government information is huge, diverse and volatile and much will be missed if memory organizations rely exclusively on approaches that selectively capture individual archival units; and (2) subject expertise and staff time to engage in such activities is increasingly scarce and expensive.

Defining authenticity

The guarantee that a digital government entity is authentic is a concern across the board. However, there is no general agreement regarding how to capture and assure the authenticity of the entity. Some respondents view the unstructured capture and re-presentation of content outside of the context of the agency web-site as increasing the risks regarding interpretation of the material (currency, authenticity, etc).

Metadata challenges

Metadata development is generally viewed as one of the most expensive and time-consuming aspects of the problem. Automated solutions are under development but in most instances the manual aspects of this activity predominate. Some suggest that the adoption of standards by government agencies would be helpful toward developing automated metadata processes (this is required by Texas statute and is a part of the implementation plan for the Texas Electronic Depository). However, preservation of web-based government information content cannot wait for legislative action. The history of the US Federal Depository Library Program reveals that even statutory obligations are difficult to enforce uniformly across a broad range of agencies.

Diverse community interests

Several of the respondents representing agency perspectives reported on the challenge of managing, presenting and preserving content in order to serve a large and diverse range of community interests. Managing and building the most appropriate formats and presentation is a serious challenge for these agencies. This challenge can certainly be extended to memory organizations that serve diverse communities.

A middle-ground or theory-neutral approach

Several respondents pointed to the need for middle-ground approaches that combined elements of bulk or targeted capture strategies, automated and manual metadata generation strategies, as well as presentation strategies that might not preserve the full "look and feel" of the original site but a convey a sufficient representation of the original content to suit local user needs. Pragmatic approaches to rights management challenges were clearly desired and practiced by several of those interviewed. Different managers and institutions clearly had differing goals and mixes of

these approaches in mind as they approached their projects. However the lack of a set of highly-functional, well-developed and easy to use set of tools with sufficient flexibility to accommodate legitimate and diverse needs of collecting programs at various memory organizations is clearly hampering progress on these fronts. [Appendix A]

Unique characteristics of web-based government content

Survey respondents indicated time and again that issues related to web-based government information are unique in certain respects (legal significance of authenticity) and that the genre in general reflects a predominance of certain types of digital curation challenges (for volatility of content and frequency of edition changes are two). Analysis of the crawl conducted for CDL under the terms of the Mellon grant will allow us to provide further insight into this question. However, the movement from project to program and incorporation of this genre within diverse digital collection programs does not suggest a unique, government documents-specific response to these challenges. For example, a recent report by UKOLN under contract with JISC reveals quite parallel findings in the domain of medical literature.²

In the next three sections we will focus on some of the discoveries we made as a result of inquiry into the capture, archives management and preservation aspects of specific projects.

Capture

We use the term “capture” to refer to various technical, administrative and intellectual activities involved in the acquisition of web-based content by memory organization, including discovery, review, selection, acquisition, etc. The various projects we surveyed are employing a wide variety of tools and practices as they approach the capture component of their projects. The issues fall into five clear areas:

Content demographics and review

There are a number of challenges associated with the content of government web sites, particularly in terms of diversity, definition and size. Webpage boundaries are often ambiguous, with linkages directly to external content and quasi-governmental sites. Almost all agency sites contain some interactive pages which allow structured access to database content. A recent study indicates that this “deep-web” content characterizes the dot gov domain at a more profound level than that of other domains: government and public sites accounted for 89.9 percent of the deep web and the National Climatic Data Center site alone accounting for 49 percent of the deep web.³ Analyses of the CDL-sponsored crawl will focus on confirming these attributes of government sites as well as other elements such as volatility of sites and frequency of new editions. It should be noted that there are as of yet no easy to use ways in which content selection specialists or bibliographers can systematically review a website to determine size, format distribution, scale of surface and deep-web or range of subjects covered by the site. Metric analysis of the sort we are conducting is subject to the constant changes of a dynamic web.

² Collecting and Preserving the World Wide Web: A Feasibility Study Conducted for JISC and the Wellcome Trust http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf

³ Michael Bergman, The Deep Web: Surfacing Hidden Value. <http://www.brightplanet.com/deepcontent/tutorials/DeepWeb/>

Selection

The various projects we reviewed range in their approaches to selection. In some of the projects, the approach has focused on “bulk collecting” as the most efficient and economic approach to acquiring content.⁴ A clear example of this approach was the *Texas Electronic Depository*, which based its approach on the notion (supported by state statutory law) that all the web-published content of Texas agencies is worthy of preservation. However, most projects have taken a more selective approach. Perhaps the most finely articulated selection policy is that of *Pandora*⁵. Individual digital government publications are captured as part of this project if they fall within a very clear and detailed set of selection based on a set of filtering criteria including subject (topic: Australia), format/medium options (preferable archival formats such as print or microform do not exist), and practical considerations (exclude dynamic pages and databases). Similarly, the *Our Digital Island Project* has developed a refined set of selection criteria for which the archival unit is the website rather than an individual publication.⁶ The Tasmanian project provides a rich set of definitions and criteria for any project focused on curation of content at the website-level including four levels of collecting ranging from comprehensive, selective, representative and snapshot:

“...*comprehensive* coverage entails capture of all updates or Webpages published in a selected Website, with the depth of coverage extending to all internal Webpages and the scope of coverage extending to primary, secondary and tertiary external Webpages...*selective* coverage entails capture of key updates or Webpages published within a selected Website, with the depth of coverage extending to all internal Webpages and scope of coverage limited to significant primary and secondary external Webpages;...*representative* coverage entails capture of occasional updates or individual Webpages published in a selected Website, with the depth of coverage limited to significant internal Webpages and scope of coverage restricted key primary external Webpages...*snapshot* coverage entails capture of individual Webpages in depth and scope sufficient only to provide a sample of the Website.” [emphases added]

Other projects focus on qualities that reflect the risk of content loss. For example, the sole selection criteria of the *Cybercemetery* is that the Federal agency producing the website is no longer in existence. At-risk content is also a contributing factor in the management of the UCLA campaign literature project; however this content is driven by the UCLA library’s existing collecting policies related to the collecting of local political campaign ephemera. *Minerva* is a very complex library involving at this point six discrete subcollections with a variety of selection principles. Two of the *Minerva* collections, the September 11 and Election 2000 collections have published selection criteria on the web.⁷ A broader selection policy for Internet content is being developed at the Library of Congress.

⁴ For a good review of the issues involved in bulk as opposed to selective collection approaches, see section 3 of the Web Preservation Project Interim Report: A report to the Library of Congress, William Y. Arms, Cornell University, January 15, 2001 <http://www.loc.gov/minerva/webpresi.pdf>

⁵ Guidelines for the Selection of Online Australian Publications Intended for Preservation by the National Library of Australia. <http://pandora.nla.gov.au/selectionguidelines.html>

⁶ Guidelines for Selecting, Archiving and Preserving Websites Pertinent to Tasmanian Government Information and Cultural Heritage. <http://odi.statelibrary.tas.gov.au/About/selpolicy.asp>

⁷ September 11 archive.org Selection Criteria <http://www.loc.gov/minerva/collect/sept11/select.html>; Election 2000 Selection Criteria <http://www.loc.gov/minerva/collect/elec2000/select.html>

The unit of selection varies from project to project. In the cases of *Minerva*, the *Cybercemetery*, *Our Digital Island*, and UCLA campaign literature collection the selection unit is the web-page. In the case of *Pandora* and the *Texas Electronic Depository*, the selection unit is a discrete digital object within webspace, such as an image or text-file.

A key issue here is that technological capacity has seriously impacted the selection policies developed by all of these projects. Preserving the look and feel of original websites is not a goal of any of the projects except perhaps for *Minerva*. In terms of open content, *Cybercemetery* excludes certain file types such as streamed videos from their capture program. Hidden or deep-web content (such as password restricted files or content requiring user-initiated completion of webforms) is not captured by any of the projects for similar reasons.

Capture tools

A variety of harvesting tools are being employed. None seem to be satisfactory to the project managers. Of the projects using more “selective” acquisitions strategies, perhaps the most advanced system is that developed by *Pandora*, an integrated selection, cataloging and archiving system developed by the Information Technology department of the National Library Australia called PANDAS (*Pandora Digital Archiving System*).⁸ The PANDAS system is built around a commercial software product, WebObjects and this allows curatorial staff to make a variety of selection decisions including frequency of site capture, file types to capture beyond the default settings, options to exclude file types or content within specified directories, etc.

Minerva has been working closely with the Internet Archive and Alexa, and LC staff was not overly enthusiastic about their success with the crawling tools used for each of their subcollections. The *Cybercemetery* has been relying on capture technology used at GPO, Teleport Pro. However, this project is being integrated within the broader UNT digital library and that program has contracted with Index Data ApS (Denmark) to develop an integrated web harvesting, metadata and content management software environment relying heavily on open-source solutions.

Metadata

Most of the projects involve capture of some baseline metadata as part of the capture process itself. In the case of many projects, this is currently handled in large part on a manual basis. The UCLA process is illustrative:

[In terms of descriptive metadata] we identify the election, the office, candidate and/or measure, plus the ‘title of the web page (usually the html title field on the home page, but sometimes that is obviously non-meaningful and we capture the most prominent words on the home page). [In terms of Administrative metadata, we record the date of capture, the software or other capture metadata used and any notes on edits to the original (other than those done automatically by the software). Metadata is typed by hand into a small text file.

Projects are experimenting with various level of automation in the metadata capture process. *Science.gov* is a project that exclusively focuses on the capture of metadata. Although in a certain sense this is not a project that represents the spirit of this proposal, we reviewed the project because it seems to be adopting an approach to spidering web-based content metadata that could be extensible to the content itself.

⁸ See PANDAS Manual: Gathering Titles. <http://pandora.nla.gov.au/manual/pandas/gathering.html>

Intellectual property

Projects are taking differing approaches to the issue of intellectual property. Those working most closely to gain consent of data providers are those involving national libraries: *Pandora* and *Minerva*. Copyright clearance is a part of the procedures for archiving in both projects. In the case of the *Cybercemetery*, the content is public and the data provider is non-existent. In the case of the *UCLA Campaign Literature Archive*, the project found that data providers were generally non-responsive when permission to archive was requested. In the cases where they did respond, they were uniformly positive and in many cases enthusiastic, often volunteering more material. As a result, the UCLA project has adopted a “capture then tell” model.

Although US public documents and California documents are generally considered copyright-free, a significant issue for both areas is in the realm of technical reports contracted with independent firms by government agencies. The terms of these contracts often contain specific provisions with regard to intellectual property aspects of the final reports. This is a subject for further review.

Archive Management

Many of the projects were confronting the issue of movement from project to program. That is, as a small project grows in content and scope, the size of the archive drives significant management and preservation concerns and strategic thinking. We see a “mainstreaming” trend emerging. Projects involving web-based government or government-related information--such as *Cybercemetery*, the *UCLA Campaign Literature Archive*, and the *Minerva* subcollections--are merging within a broader digital library program framework in which server management and preservation strategies are addressed across diverse collections and not in the context of a single project.

Metadata

Most projects are using mixed models involving some combination of machine generated and human-generated cataloging. The emerging standard seems to be XML: this is already operational in the *Minerva* Election 2002 collection.

Ingestion

With regard to ingestion of content and metadata, there is considerable distance between ideal and reality. Automated approaches are not yet refined and it was fairly clear during several of the interviews that the ingestion process was closer to a “copy and paste” approach than a systematic, automated and routine data and metadata migration process. The clear exception here is *Pandora*. And certain programs are on the cusp of implementing integrated approaches to this process for the overall organizational digital environment (Library of Congress, University of North Texas).

Presentation/rendering

Here the projects break down we assume along lines that reflect the local information user needs. Most projects re-present the content with standard views arranged along subject, title or author (corporate or personal) lines. Of the projects in which the archival unit is the website, only one has attempted systematically to preserve the look and feel of the original: the *Cybercemetery*. Even this breaks down when certain types of content (streamed video) is involved. As yet the *Cybercemetery* staff have not confronted a database-driven website.

Preservation

Again, the theme of movement of content from project to program is apparent in several instances. Project-specific preservation plans are highly schematic and incomplete. The broader memory organization within which these individual projects are sited are in general developing some form of digital preservation program that is intended to apply to diverse content. Standards for trusted digital repositories have been developed and attempts at the local level are emerging to develop compliant programs. However, in terms of the overall development of the various projects, it is clear that the preservation component is far less well-developed than either the capture or archives management components.

Appendix A: User's Functional Needs and Tools

The following represents a suite of tools that respond to the need for a common and well-developed set of tools with sufficient flexibility to accommodate a broad range of needs for the capture, curation, and preservation of web-based government information. The tools are diverse enough to be theory and content neutral.

Administrative Tools: a suite of administrative tools to assist in determining benefits, costs, and obstacles.

- **Crawl analyzer:** a tool to gather web metrics and background information about a particular website. The tool would provide data about site that would inform administrative, technical, and selection decisions about the capture, curation, and preservation of the digital entities. For, example a skim might provide information about the diversity of file formats, the size of the files, an idea about the content, and a comparison to content already captured. With this information the potential costs, value of the content, and preservation strategy could be determined.
- **Standards and best practices:** identify and adopt standard and develop best practices based on standards.
- **Rights management:** a framework for rights management.

Capture Tools: programs and tools, which automatically traverse the web by downloading documents and following links.

- Develop web **crawlers** to capture:
 1. **Static web pages:** fixed web pages, documents that can be accessed directly via search engines, generally presented via html.
 2. **Deep web:** content that resides in searchable databases (password protected, forms, etc.)
- Develop a **diverse suite of web crawlers** :
 - **Scalable web crawl:** bulk crawling, designed to scale to tens of millions of web pages
 - **Incremental crawl:** designed to update a previous crawl, updates existing set of downloaded pages
 - **Focused crawl:** designed to gather documents within specific parameters (topic, type, etc)
 - **Customized crawl:** tailored syntax of a particular site

Curation tools: a suite of tools that assist in managing digital content.

- **Automatic metadata production:** facilitates the most efficient and effective means of automated metadata production
- **“Human touch” metadata production:** integrating human and automated processes in metadata production
- **Content Indexer:** presenting content for full-text searching and browsing.
- **Automatic classification and concept mapping:** organizing content for presentation

Preservation Tools: a suite of tools that provide a flexible approach to preservation that is based on original intent of the content and the needs of the user whether in the context of a dark or light archive.

- **Preserve the of bits:** a representation of the exact bit sequence of original item
- **Preserve the content:** the words in the text, but not the format
- **Preserve the experience:** preserve the entire website experience

Appendix B: Interview Question Set A

Question Set:

1. Who is currently working on a project that is concerned with the capture, curation, and persistent management of web-based government information? Which of these projects is worth looking at?
2. What [technical] approaches do you see out there and what challenges are derived from the different approaches for the:
 - capture,
 - organization and management, and
 - preservation of web-based government information collections?
3. Can you identify the needs and challenges associated with the capture, curation, and persistent management of web-based government information from the perspective of each of the following groups:
 - the producer,
 - the memory organization, and
 - the end-user?
4. How might we document the problem – specifically how might we describe and itemize the size and composition of the .gov domain? Are you aware of any recent analysis in this area?
5. Can you suggest any ways in which we might refocus these questions? And are there any other issues you believe that we should be addressing in this survey?

Appendix C: Project Level Interview Question Set

Question Set:

1. Background and Mission.
 - a. Can you briefly describe your project, focusing on issues such as:
 - i. Funding
 - ii. Partner roles
 - iii. Stakeholder interests
 - iv. Long-range plans?
2. Selection.
 - a. How do you select material (who determines, process, criteria)?
3. Capture.
 - a. Describe the technical aspects of your capture process, focusing on tools that work well for you.
 - b. Describe the target of your capture process (domains, specific formats, etc.)
 - c. Are you missing any content in your capture process due to insufficient technology?
 - d. Open ended—what are your biggest challenges in the capture area.
4. Metadata - Ingestion.
 - a. What are your procedures for creating descriptive metadata (for identification and retrieval purposes) as well as preservation metadata (structural, administrative, and digital access management purposes).
 - b. Describe the level of automation involved..
 - c. What happens to the captured material and associated metadata once both are ready for ingestion?
5. Management - Preservation.
 - a. How are you managing the archive?
 - b. How are you preserving the content?
 - c. Do you deselect/withdraw content? Under what circumstances?
 - d. How do you handle different versions of a document?
 - e. What are your most serious challenges in managing the archive?
6. Service and Use:
 - a. How is the content accessed by and delivered to the end user?
 - b. Are there any specialized search engines / applications involved?
7. Rights Management – Legal Aspects:
 - a. Who has access to the archive?
 - b. Are there any copyright issues regarding your content?
 - c. Are there any confidentiality/privacy issues associated with the content?

**Appendix D:
Individuals Consulted**

Survey A

George Barnum
Electronic Collections Coordinator
United States Government Printing Office

Janet Fisher, Director
Library and Research Library Division
Arizona State Library, Archives & Public
Records

Gail Hodge
Senior Information Scientist
International Information Associates
Project: Science.Gov

John Jewel
Chief of State Library Services
California State Library

Survey B

Gabriella Gray, Digital Library Projects
Coordinator
Reference and Information Services
University of California, Los Angeles
Project: Online Campaign Literature Archive

Abbie Grotke
Library of Congress
Project: Project Minerva

Cathy Hartman
Head, Government Publications
Digital Library Fellow
University of North Texas
Project: CyberCemetery

Gabriella Gray, Digital Library Projects
Coordinator
Reference and Information Services
University of California, Los Angeles
Project: Online Campaign Literature Archive

Kevin Marsh
Developer, Networked Services
Texas State Library & Archives Commission
Project: Texas Electronic Depository

Richard Pearce-Moses, Director
Digital Government Information
Arizona State Library, Archives & Public
Records
Project: Web Documents Digital Archive Pilot

Margaret Phillipps, Manager of Digital
Archiving
National Library of Australia
Project: Pandora

**Appendix E:
Related Projects and Programs**
(revised 03/12/03)

United States

Federal

1. USDA Economics and Statistics System
 - a. <http://usda.mannlib.cornell.edu/usda/usda.html>
 - b. Principals: Cornell University Mann Library in cooperation with the USDA
 - c. Initiated: unknown
 - d. Scope: digital content
 - e. Contact: Rich Allen

2. Science.Gov
 - a. http://www.dtic.mil/cendi/proj_sci_gov.html
 - b. Principal: CENDI
 - c. Initiated: 2002
 - d. Scope: metadata and selective digital content
 - e. Contact: Gail Hodge

3. Web Documents Digital Archive Pilot Project
 - a. http://www.niso.org/presentations/barnum-ppt_01_22_02/
 - b. Principals: US GPO in cooperation with OCLC
 - c. Initiated: 1999
 - d. Scope: electronic US federal documents
 - e. Contact: George Barnum

4. CyberCemetery
 - a. <http://govinfo.library.unt.edu/default.html>
 - b. Principals: University of North Texas / GPO Partnership Agreement
 - c. Initiated: 1997
 - d. Scope: websites of defunct US federal agencies
 - e. Contact: Cathy Hartman

5. Minerva, the Web Preservation Project
 - a. <http://www.loc.gov/Minerva>
 - b. Principal: Library of Congress
 - c. Initiated: 2000
 - d. Scope: 2000 presidential campaign websites
 - e. Contact: Abbie Grotke

6. DOSFAN
 - a. <http://www.uic.edu/depts/lib/documents/resources/dosfan.shtml>
 - b. Principals: University of Illinois, Chicago; US Department of State, GPO Partnership Agreement
 - c. Initiated: 1993
 - d. Scope: digital US State Dept documents
 - e. Contact: John Shuler

7. National Library of Medicine
 - a. Permanence Rating System
 - b. Principals: National Library of Medicine
 - c. Initiated:
 - d. Scope: metadata and permanence rating system for NLM web sites and electronic publications
 - e. Contact: Margaret Byrnes
 - f. Reports: Phase II Report from the Permanence Ratings Committee (<http://www.nlm.nih.gov/pubs/reports/permanence.pdf>); Factsheet on preservation/Electronic Resources <http://www.nlm.nih.gov/pubs/factsheets/preservation.html>

8. USDA Digital Publications Preservation Program
 - a. <http://www.nal.usda.gov/preserve>
 - b. Principals: National Agricultural Library and the USDA Economic Research Service
 - c. Initiated:
 - d. Scope: Metadata, metadata template, framework document for the preservation of USDA digital publications (this may or may not classify as government web sites)
 - e. Contact: Evelyn Frangakis
 - f. Framework Document and other reports available via the web site

9. National Technical Information Service
 - a. NTIS Science Portals Program
 - b. Principals: National Technical Information Service/Department of Commerce
 - c. Initiated:
 - d. Scope: harvesting documents from web sites of several science agencies including Department of Energy and archiving for distribution to the public
 - e. Contact: Wally Finch
 - f. Presentation:: <http://www.science.gov/workshop/wfinch.pdf>

10. Electronic Records Archive
 - a. http://www.archives.gov/electronic_records_archives/index.html
 - b. Principals: NARA, San Diego Supercomputer Center
 - c. Initiated:
 - d. Scope: electronic records, which may include web sites; working on issues of long term preservation of various formats including pdf, various image formats, HTML, etc.; dealing with issues of collections versus more items within a collection; looking at new partnerships with agencies to share the responsibility and develop more advanced tools for support
 - e. Contact: Ken Thibodeau
 - f. Several presentations are available via the web

11. NASA Goddard Space Flight Center
 - a. <http://www.library.gsfc.nasa.gov>
 - b. Principals: NASA GSFC Library and certain GSFC codes
 - c. Scope: web sites, videos, images, project documents (some of which is made web accessible through document management systems) within the NASA GSFC domain; we are just starting and are still in a research phase to determine the most efficient way to capture and store the GSFC web pages

- d. Contact: Janet Ormes

State and Local:

- 12. Texas Electronic Depository
 - a. <http://www.tsl.state.tx.us/lot/electronicdepositorylib.html>
 - b. Scope: electronic documents
 - c. Contact: Kevin Marsh
- 13. Preserving Electronic Publications (PEP)
 - a. <http://www.isrl.uiuc.edu/pep/>
 - b. Principals: Illinois State Library, the State Library of Ohio, the Illinois Archives, and the Graduate School of Library and Information Science (GSLIS) at the University of Illinois, Urbana-Champaign. Funded by IMLS National Leadership Grant Program
 - c. Initiated: 2002
 - d. Scope: Illinois state agency webpages
 - e. Contact: Larry Jackson
 - f. Report: http://www.isrl.uiuc.edu/pep/papers/UIUCLIS_2001_9_EARCH.html
- 14. Washington state initiatives
 - a. <http://www.computerworld.com/databasetopics/data/story/0,10801,72096,00.html>
 - b. Principals: unknown
 - c. Initiated: 2002
 - d. Scope: digital government records
 - e. Contact: Jerry Handfield
- 15. Web Documents Digital Archive Pilot Project
 - a. http://www.access.gpo.gov/su_docs/fdlp/pubs/proceedings/01pro14.html
 - b. Principals: Arizona State Library, Archives and Public Records in conjunction with OCLC Web Preservation Project
 - c. Initiated: 2001
 - d. Scope: web-only state publications and records
 - e. Contact: Richard Pearce-Moses
- 16. Joint Electronic Records Repository Initiative
 - a. <http://www.ohiojunction.net/jerri/>
 - b. Principals: State Library of Ohio, the Ohio Historical Society, the Ohio Supercomputing Center, and the State of Ohio Department of Administrative Services in conjunction with OCLC's digital collection management and preservation project
 - c. Initiated: 2001
 - d. Scope: electronic public records
 - e. Contact: Charles Arp
- 17. Online Campaign Literature Archive
 - a. <http://www.library.ucla.edu/libraries/mgi/campaign/>
 - b. Principal: UCLA Library
 - c. Initiated: 2000

- d. Scope: Los Angeles campaign websites and retrospective digitization of campaign literature
 - e. Contact: Gabriella Gray
18. California Initiatives and Propositions Database
- a. <http://holmes.uchastings.edu/>
 - b. Principal: UC Hastings College of the Law with LSTA funding.
 - c. Initiated: 1999
 - d. Scope: California state initiatives and propositions, including ancillary material
 - e. Contacts: tbd

Foreign:

19. Pandora

- a. <http://pandora.nla.gov.au/index.html>
- b. Principal: National Library of Australia
- c. Initiated: June 1996
- d. Scope: government and non-governmental digital content
- e. Contact: Margaret Phillips

20. Our Digital Island

- a. <http://odi.statelibrary.tas.gov.au>
- b. Principal: State Library of Tasmania, New Zealand
- c. Scope: government and non-governmental websites
- d. Contact: tbd

21. Kulturaw3

- a. <http://www.kb.se/kw3>
- b. Principal: National Library of Sweden
- c. Initiated: 1996
- d. Scope: government websites
- e. Contact: tbd

22. Archipol

- a. <http://www.archipol.nl/english/index.html>
- b. Principal: Documentation Centre for Dutch Political Parties and the Groningen University Library, funded by the SURF foundation
<<http://www.surf.nl/en/home/index.php>>
- c. Initiated: 2000
- d. Scope: Dutch political party websites
- e. Contact: tbd

23. Nordic Web Archive

- a. <http://nwa.nb.no>
- b. Principals: each Nordic country's National Library and a Project Manager at the National Library of Norway. The project was funded by [Nordunet2](#) and the National Libraries of each Nordic country.
- c. Scope: tools
- d. Contact: tbd

24. Digitale Archivering in Vlaamse Instellingen en Diensten (DAVID)

- a. <http://www.antwerpen.be/david/eng/index.htm>
- b. Principals: Max Wildiers Foundation and is a cooperation between the Antwerp City Archives and the Interdisciplinary Centre for Law and Informatics of the K.U.Leuven.
- c. Scope: clearinghouse on digital preservation
- d. Contact: tbd

25. Netachive

- a. <http://www.netarkivet.dk/index-en.htm>
- b. Principals:
- c. Scope: tools and collecting strategies
- d. Contact: tbd

26. WARP

- a. <http://warp.ndl.go.jp/>
- b. Principals:
- c. Scope: tools
- d. Contact: tbd

Memory Organization:

27. Infomine

- a. <http://infomine.ucr.edu>
- b. Principal: University of California, Riverside
- c. Initiated: 1993
- d. Scope: metadata
- e. Contact: Keith Humphreys

Not-for-Profit Organization:

28. Internet Archive

- a. <http://www.archive.org>
- b. Principal: Brewster Kahle
- c. Initiated: 1996
- d. Scope: web-based content
- e. Contact: Michelle Kimpton

29. OCLC Web Document Digital Archive Project

- a. <http://www.oclc.org/digitalpreservation/archiving/wdda.shtm>
- b. Principals: OCLC and partner organizations
- c. Initiated: 2000
- d. Scope: web-based documents
- e. Contact: Pam Kircher (pam_kircher@oclc.org)

For-Profit Organization:

30. Google

- a. <http://www.google.com>
- b. Scope:
- c. Contact: tbd

31. Newsbank
 - a. <http://www.newsbank.com>
 - b. Scope: metadata and content
 - c. Contact: tbd

32. Lexis-Nexis/Reed-Elsevier
 - a. <http://www.reedelsevier.com>
 - b. Scope: metadata and content
 - c. Contact: tbd

33. Rand California
 - a. <http://ca.rand.org/stats/statistics.html>
 - b. Scope: statistics
 - c. Contact: Joe Nation

Appendix F: Web-based Government Information Project: Frequently Asked Questions

What is the Web-based Government Information Project?

The California Digital Library is conducting a cost-benefit review of technologies appropriate for the capture, curation, and persistent management of web-based documents of US state and federal governments. Project activities include:

- Analyze the scope of the domain of web-based government information
- Identify promising technologies and projects related to capture, management and preservation of this material
- Capture a subset of web documents from government sites to serve as a testbed for analyzing the technical requirements for building collections of web-based government documents

What is a “Memory Organization”?

A *memory organization* is any institution whose mission is to preserve the cultural legacy of societies for future generation. Libraries, museum and archives are examples of memory organization that have traditionally focused upon the preservation of content created by among others, writers, artists, scientists, scholars, government agencies and commercial organizations.

What is “Curation”?

Curation is an activity that is core to the mission of memory organizations. The OED defines the verb curate thusly: “To act as curator of (a museum, exhibits, etc.); to look after and preserve”. Curatorship is often used within the library community to describe professional staff involved in the long-term guardianship and preservation of rare books and manuscripts.

What is “Web-based Government Information”?

The focus of this project is content that is produced by government agencies or government officials and published exclusively on the web or made accessible via web-accessible interfaces. Although the project is initially on preserving static content, it is also interested in the preservation challenges posed by dynamic databases.

Isn't this work the responsibility of government agencies?

It would be wonderful if all government agencies had the resources and conscience to retain and preserve their entire intellectual output in perpetuity. In reality, few agencies possess the means or inclination to do this.

Isn't this work the responsibility of government archives and libraries?

Yes. The mission of government archives and libraries generally include the preservation of cultural records, documents and publications. However, these organizations face the same resource constraints as all government agencies. Memory organizations outside of

government have traditionally played a critical and often leading role in the preservation of cultural content in the era of print.

How does this project address the issue of authenticity of web-based government information?

Emerging technologies and metadata strategies can be adopted to ensure then authenticity of web-based government information irrespective of the server on which it resides. In addition, non-governmental memory organizations have historically performed a role as trusted repository for government information, by virtue of its independence of political imperatives and management. The volatility of digital content suggests that memory organizations outside of government should assume a leading role to ensure that government records, documents and publications are not manipulated or withdrawn to achieve politically-motivated objectives.

Who is involved in this project?

This project is a highly collaborative project supported by funds from the Andrew W. Mellon Foundation. The lead organization is the California Digital Library, where analysis of the government information domain and review of key projects are being conducted. The development of a testbed collection of web-based government documents involves collaboration with the faculty and staff at both the Stanford Computer Science Department and the San Diego Supercomputer Center.

What agency websites have been crawled in order to create the project testbed?

Six sites have been crawled including: United States Department of State; United States Geological Survey; United States Senate; California State Water Resources Control Board; California Energy Commission; and the California Legislative Analyst's Office.

What have you accomplished thus far?

Surveys of several leaders in the field as well as managers of over 25 projects are being conducted. A report summarizing the results of this work will be available by early March. Two crawls have been conducted and the contest is being ingested into a repository. A report on web metrics of web-based government information based upon a literature review and analysis of the crawl will be available in early March. A technical review of available technologies to support capture, curation and preservation of the content is underway.

What are you planning to do next?

Project staff will be presenting their reports to meetings of various stakeholder groups in the coming months, including sets of librarians who serve as intermediaries in identifying and retrieving government content, as well as representatives from key user groups including faculty, researchers, and graduate students. The goal of these meetings will be the development of a set of functional specifications for the technical infrastructure necessary to capture this genre of content as well as specification for providing a service layer surrounding this infrastructure.

Appendix G: Crawl Report and .gov demographics

Working in collaboration with The Stanford Digital Library Technologies we conducted a crawl of a handful of federal and California State sites. One of our major goals of these activities is to evaluate these technologies on a testbed system. Once the crawl has been completed the content is ported to the San Diego Supercomputer Center, SDSC.

.gov Sites Crawled:

1. U.S. Department of State and related Bureaus
2. U.S. Department of the Interior and related Bureaus
3. United States Senate
4. United States Environmental Protection Agency
5. California Energy Commission
6. California State Water Resources Control Board

Crawl Demographics:

The crawl began Nov. 25th and took approximately one week to complete. The crawl pulled in every file (including images, audio, (in theory) video, etc.). Each crawler dove down to depth 10. Stanford's current implementation either ignored or truncated files over 2MB. They are changing this.

- Using Stanford's .gov seed list, they completed 4934
- The material took about 325GB
- This corresponds to 10,668,090 files that were captured
- In comparison, a complete crawl of Stanford, pulling in all file types takes about 34GB"

Questions to ask about the crawl:

1. What is the size of the crawl? By domain?
2. What are the different file types?
3. What is the rate of occurrence of the file types
4. What is the rate of occurrence of the file types by domain (for example, does one domain produce a ton of excel files?)
5. What are the sizes of the files -- average, median, something else? What is the average size of the entity?
6. Is there any metadata that suggests when the files were originally created?
7. What is the periodicity of the data?
8. What is the % of error rate (server not responding, etc.)
9. What is rate of duplication: how many unique links, what is the rate of duplication of links?
10. How much is restricted (surface v/s deep web)?
11. Do you have an idea of the use of add-on software needed to access sites?
12. How can we determine the success of the crawl? For example, some crawlers only get about 1/3. What is our success rate?