

**Web-Based Government Information: Evaluating Solutions for Capture,
Curation, and Preservation**

An Andrew W. Mellon Funded Initiative of the California Digital Library

Project Report
November 2003

This report was compiled by Patricia Cruse, Director of Digital Preservation, California Digital Library; Charles Eckman, Principal Government Documents Librarian & Head of Social Sciences Resource Group, Green Library, Stanford University; John Kunze, Preservation Technologies Architect, California Digital Library; Heather Christenson, Resource Liaison Coordinator, California Digital Library; and with the assistance from Jennifer Colvin, Associate Editor, California Digital Library; Cate Hutton, Director of Business Development, California Digital Library; Daniel Greenstein, University Librarian and Executive Director, California Digital Library.

The **California Digital Library** < <http://www.cdlib.org>> is the University of California's 11th University library. It was established in 1997 by UC President Emeritus Richard Atkinson to build the University's digital library, assist campus libraries with sharing their resources and holdings more effectively, and provide leadership in applying technology to the development of library collections and services.

Organizationally housed at the UC Office of the President, the CDL operates in close collaboration with all UC campuses and their libraries.

Harnessing technology and innovation, and leveraging the intellectual and cultural resources of the UC, the CDL supports the assembly and creative use of the world's scholarship and knowledge for the UC libraries and the communities they serve.

To achieve its mission, the CDL utilizes strategic partnerships and technical innovation to:

- Focus on digital library collections by investing in the means of developing, acquiring access to, and persistently managing digital collections.
- Invest in applications that help campus libraries and others build meaningful, cost effective, and customized services for the use of the University's scholarly, cultural, and learning resources.
- Provide collaborative leadership in the development of new forms of scholarly communication.
- Leverage its investments to influence the marketplace for scholarly information in the interest of the University, its libraries, and the scholarly and public communities they serve.

Acknowledgements:

The California Digital Library would like to thank the following individuals and organizations that helped inform this project:

- ❑ The generous funding of the Andrew W. Mellon Foundation made this investigation possible.
- ❑ The Stanford Computer Science Department and the generous assistance of Andreas Paepcke and Hector Garcia-Molina in the use of WebBase.
- ❑ The San Diego Supercomputer Center and the generous assistance of Reagan Moore, Richard Marciano, and Charles Cowart in assisting in the use of the Storage Resource Broker and in analyzing crawl data.
- ❑ The Stanford Libraries and the invaluable contributions that Chuck Eckman made to the project.
- ❑ The survey respondents for their time and insight in helping us understand the many issues surrounding web-archiving – including: George Barnum, Electronic Collections Coordinator, United States Government Printing Office; Janet Fisher, Director Library and Research Library Division, Arizona State Library, Archives & Public Records; Gabriella Gray, Digital Library Projects Coordinator, Reference and Information Services, University of California, Los Angeles; Abbie Grotke, Library of Congress; Cathy Hartman, Head, Government Publications, Digital Library Fellow, University of North Texas; Gail Hodge, Senior Information Scientist International Information Associates; John Jewel, Chief of State Library Services California State Library; Kevin Marsh, Developer, Networked Services, Texas State Library & Archives Commission; Richard Pearce-Moses, Director, Digital Government Information Arizona State Library, Archives & Public Records; Margaret Phillips, Manager of Digital Archiving National Library of Australia.
- ❑ The Librarians that attended a one-day meeting in San Diego and brought their unique perspectives about the challenges of web-archiving – including: Jenifer Abramson, UC Los Angeles; Elizabeth Cowell, UC San Diego; Karen Fung, Stanford University (CRL Mellon project); Gabriella Gray, UC Los Angeles; James R. Jacobs, UC San Diego; Jim Jacobs, UC San Diego; Kris Kasianovitz, UC Los Angeles; Linda Kennedy, UC Davis, Mary Larsgaard, UC Santa Barbara, Carolyn Palaima, UT Austin (CRL Mellon project); Richard Pearce-Moses, Arizona State Library, Archives and Public Records; Jenny Reiswig, UC San Diego; Brad Westbrook, UC San Diego.
- ❑ The Contributions of Bernard J. Reilly, President, The Center for Research Libraries, and the insights willingly shared by CRL's Political Communication Web Archiving Project.
- ❑ The University of California government information librarians.

1. AIMS AND METHODS.....	3
1.1 <i>WEB-BASED GOVERNMENT INFORMATION: A CRITICAL AT-RISK RESOURCE</i>	3
<i>METHODOLOGY</i>	4
2 DEMOGRAPHIC REVIEW OF THE DOT-GOV DOMAIN	5
2.1 LITERATURE REVIEW	5
2.2 ANALYSIS OF THE PROJECT CRAWL	9
3 CHALLENGES IN THE CAPTURE, CURATION, AND PERSISTENT MANAGEMENT OF WEB-BASED MATERIALS.....	14
3.1 CAPTURE CHALLENGES	14
3.2 <i>CURATORIAL CHALLENGES</i>	19
3.3 <i>PRESERVATION CHALLENGES</i>	23
3.4 <i>INSTITUTIONAL READINESS</i>	27
4 ARCHIVING THE WEB: A “LAYERED” SERVICE MODEL	30
4.1 MAPPING ROLES AND INCENTIVES TO ORGANIZATIONS	32
5 A ROUTE MAP FOR SERVICE IMPLEMENTATION.....	37
5.1 REPOSITORY INFRASTRUCTURE.....	39
5.2 ENTRY POINT URL (EPU) REGISTRY	40
5.3 CURATOR INTERFACE.....	41
5.4 REGISTRY CONTENT SEARCH.....	43
5.5 ARCHIVE CONTENT SEARCH	44
5.6 CRAWLER	45
5.7 INDEXER	49
6 SUSTAINING THE BROKER SERVICE	50
6.1 COST ELEMENTS.....	50
6.2 REVENUE STREAMS	52
7 WEB ARCHIVING IN CONTEXT	55
8 CONCLUSION	60
APPENDIX 1: QUESTION SETS AND INDIVIDUALS CONSULTED	61
<i>SURVEY A</i>	63
<i>SURVEY B</i>	63
APPENDIX 2: PROJECTS AND PROGRAMS THAT INFORMED OUR RESEARCH. 65	
APPENDIX 3: SAMPLE MODELS FOR CAPTURE, CURATION, PRESERVATION .. 71	
APPENDIX 4: LIST OF SOURCES FOR SERVICE VISION	74

1. Aims and methods

1.1 Web-Based Government Information: A Critical At-Risk Resource

Government information plays a fundamental role in our society—it is a basic foundation of democracy. The data are as diverse as the agencies that create it, ranging from the Department of Health and Human Services and the National Oceanic and Atmospheric Administration, to the California State Coastal Conservancy and the California Trade and Commerce Agency. It is inherently multi-disciplinary in its appeal. And it serves multiple audiences, including research institutions, business enterprises, and private citizens.

Government information is also culturally significant. Memory organizations, government agencies, legal entities, and society as a whole rely on its existence for a variety of essential civic, economic, and political functions. These groups rely on all types of government publications, regardless of format. Digitally published materials are more volatile, uncontrolled, and at much greater risk of being lost than those that are published in printed formats. Unlike the printed publications of U.S. governments, digital ones do not flow through central printing offices, making their existence, number, provenance, and orientation impossible to record. The most volatile and at-risk government information is that which is made available exclusively via the World Wide Web, where 65 percent of all government publications that are distributed by the Government Printing Office, the largest producer of government information, are now placed without printed analog.¹

Memory organizations, particularly libraries, have a central role to play in preserving this web-based content. Selected state, public, and academic libraries already fulfill the preservation role with printed government materials, as represented by the nearly 1300 libraries nationwide who participant in the Federal Depository Library Program. These libraries maintain collections that extend back to the origins of American governments, and are uniquely positioned to ensure continuity in this historic record as government materials transition from print to digital formats. In addition, memory organizations have long-standing experience in managing and supporting the use of comprehensive collections, and in organizing those collections to meet their users' diverse needs. Additionally, memory organizations are almost unique amongst those few agencies now asserting themselves as the guardians of governments' web-based outputs.

In large part, our research set out to determine how best to leverage memory organizations' historic roles and current activities; to ask, in effect, what might encourage and enable their greater involvement in the capture and persistent management of web-based government information so that they may extend their historic roles as guarantors of our governments' copious published record into a digital age.

¹ In comments at the Federal Depository Library Council meeting in April 2003, Bruce James, Public Printer of the United States noted that nearly two-thirds of the information resources in the Federal Depository Library Program are now available only electronically. See Adler, Prudence S. Rethinking the Federal Depository Library Program. *ARL*, no. 229 (August 2003): 8. <<http://www.arl.org/newsltr/229/fdlp.html>>.

Methodology

In addition to extensive literature review and analysis, our work was conducted via a variety of means including the following activities listed below.

Demographic review of the dot-gov domain

In order to define more precisely the extent and nature of the problem that surrounds the persistent management of web-based government materials, we set out to analyze the scope, composition, and rate of change inherent in the domains of U.S. state and federal governments. The analysis was conducted in two parts: first, through literature review, data analysis, and interviews with researchers; and second, by capturing and analyzing web-based materials produced by selected U.S. national and California state government agencies. The second part of the project was conducted in conjunction with the WebBase project at Stanford University's Computer Science Department, and deployed the suite of tools that WebBase has developed to comprehensively crawl and analyze selected U.S. federal and state government agency web domains.

Surveying existing web-archiving initiatives

Here we sought to learn from existing initiatives about the key challenges that confront those who set out to build persistent collections of web-based government materials. We also sought insight into some of the more promising practices and avenues of inquiry. Information was gathered on a variety of fronts, beginning with two sets of interviews. The first set of interviews were conducted with leaders in the digital preservation and government information fields to help shape the nature of our inquiry. The second set of interviews included individuals that are directly associated with projects addressing some aspect of the problem. The survey questions and list of participants are included in *Appendix 1*.

In these interviews, we focused on three key aspects of the preservation process: capture, curation, and persistent management. We were guided by a common set of questions:

- ❑ What is the extent and nature of the problem?
- ❑ How do the needs of institutions, libraries, and memory organizations shape requirements of any solution?
- ❑ What promising solutions exist, however partial they may be? How well and at what cost do they meet the identified requirements? How might such solutions be further developed to better meet requirements? How, with what funding, and on what organizational basis might they be implemented and sustained?

We found in our interviews that there was limited experience and understanding of curation and preservation in the context of web archiving. In order to gain a comprehensive understanding of all of the issues associated with web archiving, we examined publicly available documentation of several web archiving initiatives. For a complete list of projects that we consulted see *Appendix 2*.

Surveying the needs of librarians and others who mediate between individual end users and collections of government information

We sponsored a one-day meeting of librarians from a wide range of memory organizations who had a range of collection building responsibilities in the sciences, social sciences, humanities,

area studies, and government documents. Here our goal was to gain an understanding of the challenges of web archiving from the perspective of information professionals who work on the front lines building and supporting the use of government information collections.

Iterative design and review of a web archiving service scenario

Based on these analyses, we set out to design a service scenario that promises to facilitate sustainable, cost-effective capture, curation, and persistent management of web-based materials. We worked intensively with those already active in web archiving projects, with the librarians interested in collections of government information, and with end users from a variety of scholarly disciplines. The service scenario was vetted with our informants. CDL staff developed a route map for its implementation and financial sustainability.

Setting web archiving in context

As indicated above, as much as 65 percent of all government publications that are distributed to libraries through the federal depository library program are currently produced exclusively in electronic form and distributed via the web. Still, the vast historic record of U.S. federal, state, and local governments exists in printed and other analog formats. A sizeable portion also exists on digital media that are rapidly becoming outdated. Accordingly, web archiving strategies, however comprehensive they may be, are only a partial solution to the more general problem of ensuring persistent access to all types of government information. With this in mind, we assessed some of the general challenges of the persistent management of government information.

2 Demographic review of the dot-gov domain

This analysis comprised two elements. The first involved a review of the domain through a combination of literature review, secondary data analysis, and interviews with researchers. The second involved an analysis of our test-bed crawl a selected set of government agency web domains.

2.1 Literature review

In general, the domain of web-based government information is hard to define, constantly expanding, and highly volatile. In addition, web-based government information is generally marked by a high degree of format diversity (ranging from text, to images, to databases), genre diversity (including publications, documents, and databases) and opacity (a high percentage of the content is hidden within the deep web). These and other leading characteristics are reviewed in greater detail below.

Defining the dot-gov domain. The U.S. General Services Administration has provided legal definition of the dot-gov domain and has detailed who may use it. "What is Internet GOV Domain?" the text of a federal regulation asks.

Internet GOV Domain refers to the Internet top-level domain ``dot-gov" operated by the General Services Administration for the registration of U.S. government-related domain names. In general, these names reflect the organization names in the Federal Government

and non-Federal government entities in the United States. These names are now being used to promote government services and increase the ease of finding these services.

Who may register in the dot-gov domain?

Registration in the dot-gov domain is available to official governmental organizations in the United States including Federal, State, and local governments, and Native Sovereign Nations.’²

Although the definition is clear enough in principle, it does not offer that much guidance to those interested in the persistent management of web-based government materials. Dot-gov web page boundaries are often ambiguous, with links directly to external content and quasi-governmental sites. Does the content from these pages qualify from the archivists’ or users’ perspective as government information, even though it may not qualify under the federal regulations? Many government sites do not use the dot-gov domain, or have such an ambiguous status that it isn’t clear if they are a government entity. For example, the Federal Reserve System (the Fed) serves as the nation’s central bank and is government-funded entity, yet the Fed’s web sites use the dot-org domain. Similarly, the URL for the U.S. Postal Service is www.usps.com.

Contraction and expansion. At the earliest stages of the Internet, the dot-gov domain was preeminent. However, with the growth of e-commerce and the Internet community, the government presence has declined. In 1999, Lawrence and Giles reported their analysis of the effectiveness of search engines across web domains.³ These two researchers manually analyzed a test-bed of web sites. They estimated that commercial web sites occupied 83 percent of the public web, educational domains occupied 3 percent, and government web sites reflected about 1 percent of web content. Religious, health, and other dot-org or dot-net domains occupied the remainder. An analysis of data reported by the Internet Domain Survey (presented in Table 1) reveals that the percentage of governmental U.S. hosts has steadily declined from 1992, when the dot-gov domain accounted for 9.25 percent of the overall hosts, to 2003, when it occupied approximately 0.5 percent of all hosts. On the other hand, the same data shows a 13-fold increase in the dot-gov domain from January 1992 to January 2003.⁴

² Final Rule - 41 CFR Part 102-173 - Federal Management Regulations; Internet .gov Domain (March 2003).

³ Steve Lawrence and C. Lee Giles. Accessibility and Distribution of Information on the Web. *Nature* 400:6740 (1999). <http://www.nature.com>

⁴ The Domain Survey attempts to discover every host on the Internet by doing a complete search of the Domain Name System. It is sponsored by the Internet Software Consortium, with technical operations subcontracted to [Network Wizards](http://www.isc.org/ds/). See: <http://www.isc.org/ds/>

Table 1: Hosts reported within the dot-gov domain

Analysis of Hosts Reported Within the Dot-gov Domain*				
Survey date	U.S. hosts	Dot-gov hosts	% increase of dot-gov hosts (base year=1992)	Dot-gov as a % of U.S. hosts
Jan. 1992	502572	46463	0	9.25%
Jan. 1995	4658712	175961	379%	3.78%
Jan. 1997	9797704	387280	834%	3.95%
Jan. 1999	29744280	651200	1402%	2.19%
Jan. 2001	79289850	834971	1797%	1.05%
Jan. 2003	112491240	607514	1308%	.54%

*See: <http://isc.org//ds/new-survey.html> for background information on the Internet Domains Survey.

[Note: A host is a domain name that has an IP address record associated with it. This would be any computer system connected to the Internet (via full or part-time, direct, or dialup connections) i.e. nw.com, www.nw.com]

Volatility and loss. If viewed in terms of a web of interlocking, interdependent resources, the dot-gov domain is filled with self-referencing and highly volatile content. It is characterized by broken links, as well as links to commercial and other non-governmental sites containing government-subsidized studies with proprietary content. Cho and Garcia-Molina report that the half-life of government web pages is four months⁵ – a short period, but considerably longer than the 44 day lifespan for the average web page.⁶ A forthcoming study by Lopresti and Gorin confirms that a significant percent of web-based government publications are routinely disappearing from agency pages.⁷

Format diversity. The Government Printing Office (GPO) disseminates the largest volume of U.S. government publications and information in the world. If we examine the format of the web-based materials listed in the GPO catalog, the universe of web-based government information is considerably less diverse and potentially more manageable than that represented by the dot-gov domain as a whole. An unpublished analysis by a government librarian indicates that web-based government materials listed in the GPO catalog fall into a number of common formats: 88 percent are PDF files, and the remainder are either in HTML, text or word-processed formats.⁸ Larry Jackson analyzed cataloged web-based state documents produced by Arizona and Illinois state governments. His study indicated a higher percentage of HTML and other markup formats in those state domains than in the federal domain.⁹ Unpublished data regarding the NARA Presidential Website preservation project revealed a disproportionately high percentage

⁵ Cho, J. & Garcia-Molina, H. *The Evolution Of The Web And Implications For An Incremental Crawler*. December 2, 1999.

<http://citeseer.nj.nec.com/rd/8318624%2C466777%2C1%2C0.25%2CDownload/http://citeseer.nj.nec.com/cache/papers/cs/23738/http://zSzzSzwww.vldb.org/zSszconfzSz2000zSzP200.pdf/cho00evolution.pdf>

⁶ Lyman looks at the entire web and discovers the average lifespan of a webpage is 44 days. See Lyman, Peter. Archiving the World Wide Web. In *Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving*. Washington DC: Council on Library and Information Resources, Library of Congress, April 2002. pp. 38-51.

⁷ Lopresti, Robert and Gorin, Marcia. The Availability of U.S. government depository publications on the World Wide Web. *Forthcoming in Journal of Government Information* 29:1.

⁸ Vassilakos-Long, Jill. Unpublished manuscript. California State University, San Bernardino. 2003.

⁹ Jackson, Larry. *Statistical Profiles of Web and Metadata Usage by Two U.S. State Governments*. 2002. http://www.isrl.uiuc.edu/pep/papers/UIUCLIS_2002_6ARCH/

of HTML files (56 percent) on the four presidential sites within the scope of that project, with the remainder files in GIF (15 percent) and PDF (13 percent) formats. A small percentage of files were in other formats.¹⁰ The discrepancy with the Government Printing Office may reflect the GPO's targeting of publication and document-like content.

Genre diversity. The genre of government information in the print world mostly consists of publications, documents, and records. These document categories continue to be reflected in the web-based realm. Publications are largely issued as presentation format files in PDF; in some cases they are described and summarized on the web, but continue to be issued in paper for sale. Documents tend to be issued in a wider range of file formats, including PDF, TIFF, word-processing, or HTML.

Organizational information (such as organization charts, directories, and job announcements), as well as current awareness information (such as press releases and "what's new" content) is heavily represented on the web. It is also the type of content most prone to change.

The agency "Homepage" is, of course, unique to the web; it has no print or other analog equivalent. Agencies often invest significant funds in the design and development of these pages. Homepages often provide timely information about developments at the agency, serving an invaluable function in depicting the agency's role, policy focus, and activities. Because this information is produced to be timely, it is also amongst the most highly volatile of all web-based government materials.

A second, relatively new category of information involves "services." These include the transaction-based services often referred to as "e-government." This category also includes databases and web query forms related to the retrieval of information from government databases of records such as documents, technical reports, metadata for publications, contact names, statistics, etc. Although there were some precursors to these web-based services in the mainframe dial-up world, the growth of the Internet and associated software tools has led to a dramatic growth of this category of government information.

Opacity. Many federal and state agency web sites provide a dynamic search interface layer on top of their document and technical report collections. Examples of this approach include the Department of Energy's Information Bridge database <<http://www.osti.gov/bridge/>> and the Government Printing Office's GPO Access services <<http://www.gpoaccess.gov/>>. Due to these interfaces, the bulk of federal content is effectively inaccessible to web crawlers. The scope of the problem is significant: Bergman (2001) indicates that dot-gov domain occupies more than 85 percent of the deep web.¹¹

¹⁰ Unpublished data provided by San Diego Supercomputer Center.

¹¹ A single government site, the National Climatic Data Center, accounted for 49 percent of the deep web. Bergman, Michael K. The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing* 7:1 (2001). <http://www.press.umich.edu/jep/07-01/bergman.html>

2.2 Analysis of the project crawl

Working in collaboration with Stanford Digital Library Technologies and the San Diego Supercomputer Center (SDSC), we conducted two test crawls of a selected set of federal and California state sites. Our major goal was to gather detailed data on the dot-gov domain in order to understand the domain's unique nature and the challenges of capturing, curating, and preserving web-based government information.

We worked with staff from the Stanford Digital Library Technologies project to formulate the crawl specifics. WebBase, the crawler developed by the Stanford Digital Library Technologies Project, was then started by the Stanford team. After the crawl was completed, the contents of the crawl were moved to a Storage Resource Broker (SRB) collection at the SDSC using a bulk registration tool. Staff at the SDSC worked with us to formulate a list of questions that were used to analyze the crawl and illuminate the demographics of the captured sites.

The goal of the crawl was to capture as wide a range as possible of the content types and formats used for web-based government information. Thus, we attempted to obtain data from a set of large sites, which were selected for their variety of document formats, types, and subject matters. Another broad goal was to crawl both U.S. federal government sites and California state government sites in order to compare and contrast their demographics. For the purposes of this report, data from one sample crawl was analyzed.¹²

The crawls took approximately one week to complete. The crawler was configured to capture all file types (including images, audio, video, etc.), to drill down to a depth of 10, to ignore or truncate files over 2MB, and retrieve a maximum number of just over 38,000 files from a site.¹³

The following sites were selected as starting points for the crawl:

- ❑ U.S. Department of State and related bureaus
- ❑ U.S. Department of the Interior and related bureaus
- ❑ U.S. Senate
- ❑ U.S. Environmental Protection Agency
- ❑ California Energy Commission
- ❑ California Legislative Analysts Office
- ❑ California State Water Resources Control Board

Summary of metrics of the first sample crawl:

- ❑ Total number of domains: 5,421
- ❑ Total number of files: 16,937,553
- ❑ Total size: 1,562,304,327,352 bytes
- ❑ Mean bytes per file: 92,239

Although these data comes from just one sample crawl, they are large enough to allow for some general conclusions about the composition of sites in the dot-gov domain. Configuration of the

¹² Raw data from the sample crawl are available from the California Digital Library on request.

¹³ These latter two limitations were changed in subsequent crawls.

crawler allowed a large amount of data to be collected outside our target list of agencies. In analyzing the crawl data, we were careful to distinguish analysis of data about the web's content (i.e., file type and size) from crawler performance (i.e., speed, duplication, and errors). The patterns observed are generally in agreement with findings surfaced in our literature review and are discussed in some detail below.

Web site size. The crawl data suggest that most government sites contain less than 10,000 files. Fully 92 percent of sites returned less than 10,000 files: 63 percent of sites returned less than 1,000 files, and another 29 percent returned 1,000–9,999 files. The site size metrics are represented in Figure 1.

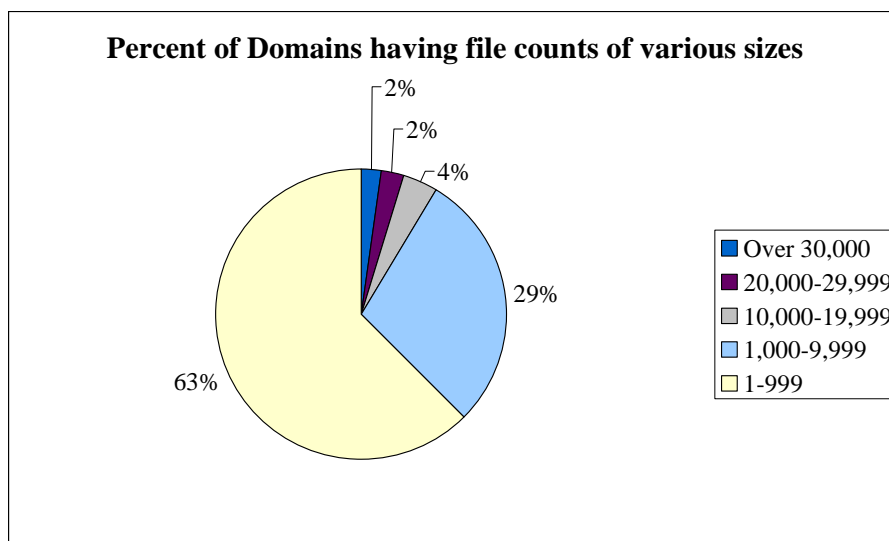


Figure 1: Percent of domains having file counts of various sizes

The largest sites in both federal and state governments (those with the largest file counts) are generally managed by the major agencies and contain large databases and voluminous educational materials (Table 2). Many produce a large amount of scientific information. The National Aeronautics and Space Administration <www.nasa.gov>, the National Oceanic and Atmospheric Administration <www.noaa.gov>, and the California State Water Resources Control Board <<http://www.swrcb.ca.gov>>, for example, produce large amounts of data, maps, and technical reports. Their web pages also have a large number of graphics.

Table 2 File count of federal and state agency sites

Sites with the Largest Number of Files: Federal		
Agency	Domain	File Count
National Aeronautics and Space Administration	http://quest.arc.nasa.gov/	38,331
Bureau of Land Management	http://www.blm.gov/	38,330
National Aeronautics and Space Administration	http://igsceb.jpl.nasa.gov/	38,330
National Oceanic and Atmospheric Administration	http://www.cpc.ncep.noaa.gov/	38,328
Department of Energy	http://envirotext.eh.doe.gov/	38,328
Sites with the Largest Number of Files: California State		
Calif. Department of Education	http://goldmine.cde.ca.gov/	38,312
Calif. State Water Resources Control Board	http://www.swrcb.ca.gov/	38,216
Calif. Air Resources Board	http://www.arb.ca.gov/	38,175
Calif. Department of Education	http://star.cde.ca.gov/	37,647
Calif. Department of Industrial Relations	http://www.dir.ca.gov/	29,715

Sites with the largest files (measured in bytes per file) generally belong to agencies that produce large amounts of scientific information—datasets, maps, and technical reports (Table 3). Further investigation reveals that these agencies also have many large PDF files. The FAA Office of Aerospace Medicine and NASA Small Aircraft Transportation System have PDF files with page counts that typically exceed 1,000 pages.

Table 3. Largest file size of federal and state agency sites (measured in bytes/file)

Largest Bytes per File: Federal Agencies		
Agency	Domain	Bytes/File
U.S. Geological Survey	http://access.usgs.gov/	20,654,308.21
National Aeronautics and Space Administration	http://denali.gsfc.nasa.gov:8001/	11,651,477.87
Mojave Desert Ecosystem Program	http://www.mojavedata.gov/	9,846,414.46
U.S. Courts	http://ca6.uscourts.gov/	4,878,681.50
National Aeronautics and Space Administration	http://modis-land.gsfc.nasa.gov/	3,732,367.11
Largest Bytes per File: California State		
Calif. Department of Energy	http://www.energy.ca.gov/	583,742.38
Calif. State Water Resources Control Board	http://www.swrcb.ca.gov/	318,531.51
Calif. Department of Water Resources Interagency Ecological Program	http://www.iiep.water.ca.gov/	309,879.21
Calif. Air Resources Board	http://www.arb.ca.gov/	116,641.07
Calif. State Board of Equalization	http://www.boe.ca.gov/	169,926.28

Format diversity. The crawl captured 335 different file types. However, the spread is uneven, with only four file types constituting 95 percent of the overall spectrum. As shown in Figure 2, HTML was by far the most common file type, representing 60 percent of files, followed by the GIF and JPEG image formats, representing 15 and 10 percent of the files, respectively. PDF was the fourth most common file type (representing 10 percent of files), but made up the highest

percentage of bytes. File types other than HTML, GIF, JPEG, and PDF constituted the remaining 5 percent of the sample. These data on format diversity also agree with the information revealed in our literature review.

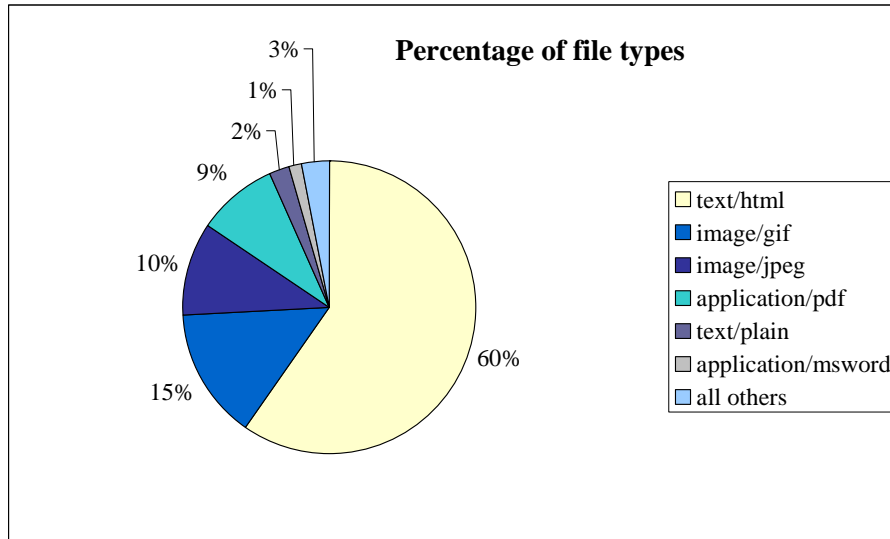


Figure 2. Distribution of file types (by file) on government and state agency web sites

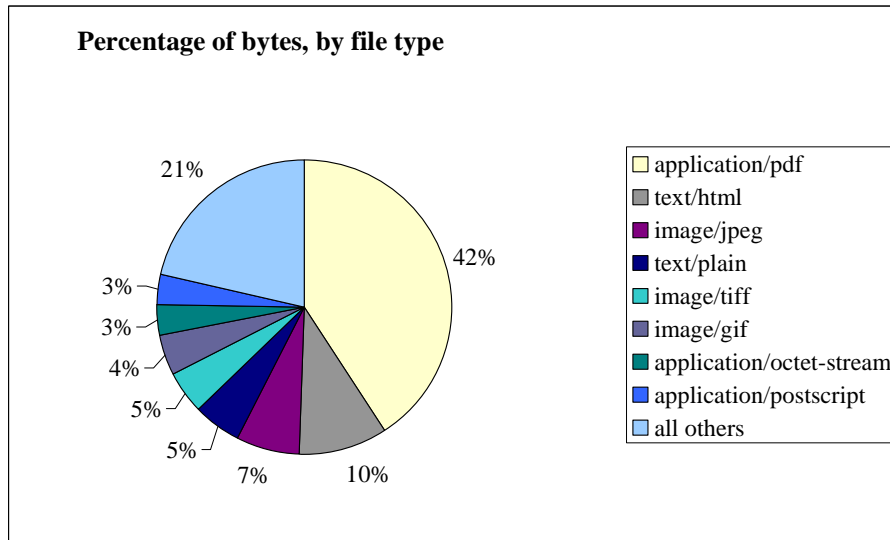


Figure 3 Distribution of file types (by byte size) on government and state agency web sites.

Errors encountered by crawl. The error percentage reported for the sample crawl was very low—only 3.68 percent (Table 4). The array of status codes for pages returned in the sample crawl supports previous findings about the volatility of the dot-gov domain. Other than the “okay” status code 200, the status codes with the highest frequency were the “found” code 302 (redirection), and the “not found” code 404. The following is a full chart of the status codes returned in the sample crawl

Table 4. Errors encountered by crawl and their frequency

Errors by code for sample crawl		Status code	Frequency
Successful	Okay	200	15854091
	No Content	204	230
Redirection	Multiple Choices	300	1322
	Moved Permanently	301	70279
	Found	302	389001
Client Errors	Bad Request	400	22675
	Unauthorized	401	28378
	Forbidden	403	13684
	Not Found	404	521165
	Method Not Allowed	405	406
	Not Acceptable	406	710
	Request Timeout	408	413
	Gone	410	412
	Length Required	411	412
	Request URI Too Long	414	436
	Server Errors	Internal Server Error	500
Not Implemented		501	534
Bad Gateway		502	1197
Service Unavailable		503	1544
Gateway Timeout		504	1433
Unknown		506	546
CGI Error		599	600
Processing Error		No Status Code Found	null

Conclusion

Crawl data such as this will be a rich source for further research in a number of areas. We intend to continue our analysis of the testbed by comparing state and federal sites. As data from our second crawl becomes available for analysis, we will also develop a set of change-monitoring metrics to track the characteristics of domain volatility. Although it may not be possible to do a file-by-file comparison of content captured over time, it should be possible to analyze changes in individual site size and file type, and size distributions within those sites.

In the meantime, our analyses to date suggest two things. First, the government domain is relatively large, volatile, and opaque, and comprises a small number of constituent file types. The relative homogeneity of the file types as highly standard image and text presentation formats suggests that the real preservation challenges relate to the size and short half-life of the information, rather than to its format diversity. Second, the demographic characteristics of the dot-gov domain are not uniformly apparent. Since the size, depth, format, and genre diversity, and rate of change of web-based materials are likely to influence data capture and preservation strategies, a comprehensive demographic analysis will be a vital step in planning a preservation program that focuses on one or many parts of dot-gov domain.

3 Challenges in the capture, curation, and persistent management of web-based materials

In this section, we review the key challenges that web-archiving projects face in the capture, curation, and persistent management of web-based materials. We also briefly discuss a further challenge that we did not anticipate—that of building institutional capacity around a web-archiving initiative that is capable of sustainably supporting it. Discussion is based upon interviews with those involved in web archiving projects for government materials and with librarians and others who have a stake in their work. Since web archiving is a new and immature area of endeavor, we found ourselves relying heavily on the extant literature in areas where our informants had little practical experience.

3.1 Capture challenges

We use the term “capture” to refer to the various technical, administrative, and intellectual activities involved in the acquisition of web-based content by memory organizations, including their discovery, review, selection, and acquisition. The various projects we surveyed are employing a wide variety of strategies and tools. The challenges they face fall into five areas: 1) selection; 2) analysis of web demographics; 3) selection of appropriate capture strategies and tools; 4) applying metadata at capture; and 5) copyright clearance.

Selection

Selection is complicated by a number of challenges. First, it is unlikely that any one definition of the government domain is ever likely to be agreed upon by those who set out to archive it. The domain is highly permeable and its boundaries are uncertain, as indicated in section 2.1 above. This boundary definition problem is fundamental to web archivists who need to scope the collections they set out to build. It is influential in another way as well. Among our respondents, there were some who were apprehensive about collecting outside the dot-gov domain because of their assumption—itself a problematic one—that everything associated with dot-gov is in the public domain and is free and clear of copyright considerations.

Second, organizations that preserve web-based government information do so under a variety of very different circumstances. They serve different audiences and exist in very different political, financial, and technical regimes. Together, these influences shape very different selection policies about what to collect. The Texas Electronic Depository grounds its collection policy in state statutory law that determines that all of the web-published content of Texas government agencies is worthy of preservation. Most of the other projects we reviewed take a more selective approach. The PANDORA project has perhaps the most highly selective approach to content selection.¹⁴ The project captures individual digital government publications if they fall within a detailed set of filtering criteria, including subject (topic: Australia), format/medium options (when preferable archival formats such as print or microform do not exist), and practical considerations (the decision to exclude dynamic pages and databases). Similarly, Our Digital Island has developed a refined set of selection criteria focusing on web sites as the archival units,

¹⁴Guidelines for the Selection of Online Australian Publications Intended for Preservation by the National Library of Australia: <http://pandora.nla.gov.au/selectionguidelines.html>

rather than on individual publications that may appear within a site.¹⁵ This Tasmanian project provides a rich set of definitions and criteria for any project focused on the curation of content at the web site level, including four levels of collecting (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 exclusively on content that is deemed most at risk for loss. For example, the CyberCemetery only collects web sites created by government agencies that no longer exist. At-risk content is also a contributing factor in the management of the UCLA Campaign Literature Archive. And here, selection decisions are further refined with reference to the UCLA library’s existing collecting policies, which emphasize the collection of local political campaign ephemera. Minerva is a complex digital library that includes six discrete collections, each of which has its own selection policy. Two of the Minerva’s collections, the September 11 and Election 2000 collections, have published selection criteria on the web, while an umbrella policy governing the Library of Congress’s acquisition of general web-based materials is under development.¹⁶ The unit of selection also varies from project to project. In the cases of Minerva, the CyberCemetery, Our Digital Island, and the UCLA Campaign Literature Archive, the selection unit is the web page. In the cases of PANDORA and the Texas Electronic Depository, the selection unit is a discrete digital object within the web space, such as an image or text file.

Third, selection decisions are influenced by the pressures to ensure the authenticity of an archive’s contents. Authenticity is a particular concern for those who persistently manage government information in any format, since it takes on legal significance. In a web environment, the rapid rate that information changes makes authenticity difficult to assess, verify, and record.¹⁷ Whether these concerns are unique among government information web archivists is not entirely clear. A report by UKOLN, a center of expertise in digital information management located at the University of Bath, under contract with JISC (Joint Information Systems Committee), reveals a similar concern experienced by those who are interested in

¹⁵ Guidelines for Selecting, Archiving and Preserving Websites Pertinent to Tasmanian Government Information and Cultural Heritage: <http://odi.statelibrary.tas.gov.au/About/selpolicy.asp>

¹⁶ 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>

¹⁷ Bearman and Trant point out that concerns about authenticity aren’t new. They are more complicated, however, in a web environment where information is replicated so extensively. See Bearman, David, and Jennifer Trant. Authenticity of Digital Resources: Towards a Statement of Requirements in the Research Process, *D-Lib Magazine*. (June) <http://www.dlib.org/dlib/june98/06bearman.html>

preserving web-based medical literature.¹⁸ In any case, there is no general agreement about how to capture and assure the authenticity of web-based information, and this uncertainty is influencing selection decisions. Some of our respondents believed the unstructured capture and re-presentation of content outside of the context of a government agency web site would increase the risks of misinterpreting the captured material (e.g., as a once current and authentic output of a government entity). Accordingly, they are steering clear of content beyond the dot-gov domain, irrespective of other reasons that might argue in favor of its inclusion in their archive.

Analysis of web demographics

The demographics of the dot-gov domain (including the nature of its content, and its rate of growth and change) encountered by a preservation initiative are directly related to the initiative's selection decisions. Accordingly, the value of government web demographic data is limited when developing effective strategies for data capture, curation, and persistent management. This is not to say that we have no reliable information about the government domain. On the contrary, we have a great deal. As already demonstrated, in the dot-gov domain, web page boundaries are often ambiguous, with links to external content and quasi-governmental sites. Almost all agency sites contain some interactive pages, which allow structured access to database content. And the dot-gov domain may have more deep web content than any other part of the web.¹⁹ The problem, according to our analysis of selected state and federal government agency web sites, is that these demographic attributes are not distributed evenly across the entire domain. For this reason, an essential step for any preservation initiative that attempts to persistently manage some part of the government web is to analyze the demographics of a particular segment in considerable detail. And yet, there are currently no tools that are readily available and relatively easy to use that can assist in this type of discreet review.

Selection of appropriate capture strategies and tools

The strategies and tools that preservation projects use to capture web-based government materials are as varied as their selection policies. Indeed, we found that decisions about selection and capture are intimately linked. Capture strategies can be said to exist along a spectrum with automated or "bulk" collecting at one end and more selective collecting at the other. A review of the pros and cons of these two approaches precedes an analysis of how they are closely associated with selection in the projects we examined.

Bulk collecting is a largely an automated method where a crawler performs a global capture without prior selection or analysis activities, other than targeting a domain or set of sites. The Internet Archive's Wayback Machine is one of the most well known initiatives using this approach.²⁰ While bulk collection is often cited for being economical,²¹ many pointed out

¹⁸Day, Michael. *Collecting and Preserving the World Wide Web: A Feasibility Study Undertaken for the JISC and Wellcome Trust*. UKOLN University of Bath, 2003.

http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf

¹⁹ For example, government and public sites accounted for 89.9 percent of the deep web; the National Climatic Data Center site alone accounted for 49 percent of the deep web. Bergman, Michael K. The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing* 7:1 (2001). <http://www.press.umich.edu/jep/07-01/bergman.html>

²⁰ The Internet Archive <www.archive.org> is building a digital library of Internet sites and other cultural artifacts in digital form.

²¹Arms, William.Y. *Web Preservation Project Interim Report: A report to the Library of Congress*. Washington, D.C.: Library of Congress, January 15, 2001. <http://www.loc.gov/minerva/webpresi.pdf>

inherent flaws in this method: (1) databases, password protected files, 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; and (3) capture may involve too many throw-away items if not carefully targeted, and targeting is not well-supported among existing technologies. Proponents of bulk collecting models are quick to point out that one person's trash is another person's treasure—that it is difficult to determine the potential and future use of captured information. Just as compelling is the often-cited argument that the low cost of disc space far outweighs any benefits of alternative approaches for collecting information.

Selective strategies involve the capture of files (by crawling or other means) that meet specific criteria. Sites are selected based on some a priori knowledge about them and about the potential for their future use. The Minerva project uses a selective capture strategy to collect web sites which have been selected by “Recommending Officers” in consultation with the MINERVA Team²² Negotiated capture represents a particular flavor of selectivity. Here the web archive actually negotiates with the producer of web-based information before securing that information in the archive. The PANDORA project uses this approach.

Several respondents clearly articulated two distinct disadvantages involved in the selective method 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; (2) subject expertise and staff time are increasingly scarce and expensive.

Choice of a capture strategy is related to, and to some extent, determined by selection decisions. For example, the Texas Electronic Depository has adopted a bulk capture strategy as the most efficient and economic way to acquire all of the web-published content of Texas government agencies.²³ PANDORA's capture strategy permits and reflects the high degree of selectivity inherent in its collection policy. The relationship between selection and capture runs in two directions. In several of the projects we reviewed, selection policies were shaped by available technological capacity. For example, only one of the projects (Minerva) sets out to preserve the look, feel, and functionality of the web sites it captures. For others, the technological barriers to this approach are too high to make it worth considering. Technological constraints also steer the CyberCemetery away from certain file types, such as streamed videos. Hidden or deep web content (such as password restricted files or content requiring the user-initiated completion of web forms) is not captured by any of the projects, again because it isn't technically feasible to capture such content.

Where capture tools are concerned, the projects we reviewed are using a great variety, and none seem to be satisfactory. Of the projects using more selective acquisition strategies, perhaps the most advanced system is that developed by PANDORA, the PANDORA Digital Archiving System (PANDAS). Developed by the Information Technology department of the National Library Australia, PANDAS is an integrated selection, cataloging, and archiving system built

²² See Minerva's Collections Policy Statement: <http://lcweb.loc.gov/acq/devpol/webarchive.html>

²³ For a good review of the issues involved in bulk and selective collection approaches, see section 3 of the report by Arms, William.Y. *Web Preservation Project Interim Report: A report to the Library of Congress*. Washington, D.C.: Library of Congress, January 15, 2001. <http://www.loc.gov/minerva/webpresi.pdf>

around WebObjects, a commercial software product. It 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 worked closely with the Internet Archive to capture a set of web sites related to the U.S. 2000 presidential election and to the September 11 tragedy. The Internet Archive, a public non-profit organization (working with Alexa Internet) collected open-access HTML pages and associated images. While the Library of Congress staff acknowledged the value of the Internet Archive's service, they also pointed out the shortcomings of the crawls. The staff felt they were not always successful in crawling each of their sub-collections. The recent report by UKOLN under contract with the JISC Wellcome Trust also found that "significant content or functionality may be missing" from Internet Archives content.²⁴

The CyberCemetery has been relying on capture technology used at the Government Printing Office, Teleport Pro. However, this project is being integrated within the broader University of North Texas digital library program, which 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.

Applying metadata at capture

Metadata development is generally viewed as one of the most expensive and time-consuming tasks in the preservation of web-based digital materials. This subject is dealt with in the section on curation below. Automated solutions for the application of metadata 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 for developing automated processes that create rich metadata for web information as it is captured. The state of Texas, for example, has implemented a statutory requirement that state agencies insert meta-tags into their state agency electronic documents. Other projects felt that their preservation efforts could not wait for similar legislative action.

Most of the projects surveyed acquire some baseline metadata as part of the data-capture process. This information—describing the object's original URL, the time of its capture, HTTP headers (content type, language, character set, file size), and the locally assigned identifier (e.g., file name)—provides the basis for archival management activities such as inventory, validation, troubleshooting, capacity planning, and reporting. Some projects are experimenting with automatic processes; most, however, handled, metadata 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

²⁴ Day, Michael. *Collecting and Preserving the World Wide Web: A Feasibility Study Undertaken for the JISC and Wellcome Trust*. UKOLN University of Bath, 2003.
http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf

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.

Any more precise characterization of the metadata a project generates (automatically or by hand) is impossible. At this level, choices are highly idiosyncratic and seem to be contingent upon selection decisions (what materials the archive is collecting), use considerations (who is expected to use the archive and for what purposes), and capture methods. In this regard, our review suggests that a range of very different approaches to metadata development are likely to flourish as organizations build persistent collections of web-based government materials to meet specific, local collection and user needs.

Copyright clearance

Projects adopt very different approaches to copyright clearance as well. These differences have less to do with legal considerations than they do with selection and capture decisions, cost-benefit analysis of effort required to clear copyright (however informal that analysis might be), and the mission orientation of the organization's web archiving project. The organization's mission orientation has the greatest impact on the PANDORA and Minerva projects. Both projects are based in national libraries that have statutory copyright obligations. Not surprisingly, perhaps, these projects work assiduously to gain consent from web site owners whose content they capture and enter into the archives.

In the case of the CyberCemetery, feasibility seems to be a determining influence. The captured content is public (insofar as it is freely available on the web) and the data provider is non-existent, so copyright clearance—highly impractical, if not impossible—is deemed to be a non-issue. UCLA's Campaign Literature Archive seeks a middle road. Initially the project sought consent from data owners, but they found owners were generally unresponsive to requests for permission to archive their content. The data providers that did respond to the project's requests were uniformly positive and in many cases enthusiastic, often volunteering more material for the archive. As a result, the UCLA project has adopted a "capture then tell" approach to copyright clearance.

Legal issues, though not at the forefront for any of the projects we reviewed, are also a consideration. Although the documents produced by U.S. federal and state governments are generally considered to be free of any copyright restrictions, there are significant issues regarding the technical reports that government agencies contract with independent firms. The terms of these contracts often contain very specific provisions about the intellectual rights and copyrights vested in reports. This is a subject for further review, but it suggests that copyright clearance procedures (and likely, access restrictions) will be required at least by some organizations that collect web-based government content.

3.2 Curatorial challenges

Curation is used in a broad sense to refer to a range of activities through which collections of web-based materials are organized, enriched, managed, and made available. While our informants have a wealth of experience in the curation of controlled and stable information

resources (e.g., traditional print and manuscript collections), our interviews show that curation techniques for web-based material are relatively unknown. Having said that, our survey revealed a number of common challenges (discussed below). Most important was the enormous variety of approaches we identified in how web-archiving projects confronted these challenges. As with approaches to capture, approaches to curation are dictated primarily by selection decisions and secondarily by the capture strategies deployed to support them. The way information is organized and made accessible reflects on the type of information and how it was acquired. Considerations about how an archival collection will be used also come into play. Again, the data suggest that a diversity of approaches is essential and will likely continue. Simply put, the archiving practices adopted by institutions for persistent management of web-based government materials are shaped by the institution's strategic objectives, its historic mission, and by the users that it serves. Technological considerations are apparent, but seem insignificant relative to others. Perhaps one reason is this: web archiving is so new that any venture into it, no matter how motivated, is intrinsically a commitment to the adoption—in some cases the development—of new technologies.

Naming, duplicate detection, and removal

Here we refer to the processes that an archive uses to name information objects as they are captured and to record their arrival in an inventory database of some kind. Duplicate detection refers to the identification (leading to possible withdrawal from the archive) of captured information objects that already exist in the archive. Removal refers to both the policies and processes for removing an object from an archival collection. The processes currently in use vary considerably, and mainly reflect the capture methods that are being deployed. Archives using automated or bulk capture techniques, for example, are relying more heavily on automated processes than more selective or manual capture techniques.

The main challenge in naming is to accurately identify each downloaded object for future internal and external reference. Archives that are highly selective in their capture (whether they are using by-hand techniques or smaller-scale web-crawlers) are typically exploiting the fact that every fully-qualified URL (containing a complete hostname and absolute pathname) is globally unique for at least a few moments before and after an object at that URL is downloaded. This provides a reasonable basis for constructing a local identifier. With as little alteration as possible, the URL is converted into a legal name in the local file system, and the object is stored in the named file. In this manner an entire remote web site can be mirrored with the original hierarchical structure preserved.

The naming challenge is considerably more complex for projects that are using automated bulk-capture techniques because of the quantity of objects that need to be dealt with. For example, using identifiers derived from each object's URL, a wide, deep hierarchy can generate such a large set of long identifiers that alternate identifier strategies have emerged. The Storage Resource Broker system flattens the file directory structure on ingest and packs multiple files into named containers before storing them, but stores the hierarchical layout in a metadata catalog (MCAT). Stanford University's WebBase crawler uses a different approach. WebBase

stores a few very large files created by concatenating the crawled file headers and content for each downloaded object.²⁵

Despite their promise, automated approaches to naming are not yet fully developed. During several of the interviews, it became clear that processes described by informants as automated are in fact closer to semi-automated "copy and paste" routines that are driven through human intervention rather than fully automated. The PANDORA project stands out as the single exception, as they are well on the way to an automated naming procedure. Further, a number of other projects (Library of Congress, University of North Texas) seem to be on the verge of implementing more fully automated acquisitions procedures.

Duplicate detection is a general problem for library and archival collections, and web archives are no exception. The process relies intrinsically on automated routines, even where archival collections are developed through highly selective or even by-hand capture mechanisms. Simply put, even the smallest and most selective collection will soon grow to a point where it becomes impossible to detect duplication with the human eye. Where automated bulk-capture processes are used, some level of duplicate removal is assured by supporting crawlers. Most crawlers will perform some duplicate removal to ensure they do not download an information object twice during the same crawl. They will not, however, compare downloaded material with what is already in the archive, so they cannot identify what information objects are being captured and stored redundantly. Here, one can rely on techniques that enable early detection of non-duplicates by comparing message digests (e.g., MD5 checksums) and file lengths. While this helps narrow the field of likely duplicates considerably, the cases that remain can only be identified as duplicates through relatively compute-intensive means. Tradeoffs in storage, development costs, and service priorities will likely lead organizations to tolerate some level of duplication within their archives.

The problem involved in detecting new or related versions of a document that already exists in the archive (near duplicates) is similar to that involved in detecting duplicates. It will be difficult to automate version detection unless versions of an object are named in such a way to indicate the precise nature of the relationship. However, even if the objects are named to indicate their relationship to one another, there is currently no standard practice that would eliminate the need for site-specific solutions.

Different archives may make different decisions about whether to retain or discard duplicates and near duplicates. Even those that are tolerant of duplication will have to anticipate removing some objects from the archive if only because curators and crawlers will make mistakes. Reasons (other than errors) for abandoning an object include assertion of copyright, revelation of fraud, and charges of libel. While these may be insufficient to warrant an object's complete removal from the darkest web archive, they can still affect its visibility in access systems and its estimated value and priority to the archiving organization. Things to consider in such cases are risks and object storage costs, as well as policies on name re-use and access to removal logs.

²⁵ See <http://www.sdsc.edu/DICE/SRB/Pappres/Pappres.html> for recent publications and presentations on SDSC's Storage Resource Broker.

Metadata enrichment

After being named (as discussed above), objects acquired for the archive are assigned metadata that helps assist in its management, discovery, and presentation. This activity can become arbitrarily complex. Most projects are using mixed models involving some combination of machine-generated and human-generated metadata assignment. The emerging standard format seems to be XML; this is already operational, for example, in the Minerva Election 2002 collection. Outside encoding format, however, there is little standards convergence. Given the immaturity of web archiving initiatives, then, at this stage it is only possible to report on some of the metadata enrichment issues that remain to be addressed.

As already discussed, the projects we reviewed generate basic metadata (about an object's original URL, the time of its capture, HTTP headers, and the locally assigned identifier) as objects are captured (regardless if they use automated or manual processes). This information provides the basis for archival management activities, including inventory, validation, troubleshooting, capacity planning, and reporting. Such metadata, however, will not support all archival management activities. For instance, potentially offensive or illegal materials that are discovered via an access system could become a political (hence a funding) liability for an archive, unless the archive is able to take steps quickly and effectively to remove the material from public view. Information to support these activities cannot easily or automatically be identified and tagged. Supplying it, therefore, remains one of the major metadata enrichment challenges.

A further challenge is the accurate identification of a captured object's format. Objects downloaded with HTTP are frequently accompanied by a content designation header that should identify the format. The header, however, is not always used. Where it is used, it is not always correct. When it is used and correct, it is often non-standard. A less-used format designation is more likely to be recognized by tools (e.g., browsers) that are contemporaneous with the object. As time goes on, we are likely to encounter more objects with old format synonyms that have long-since been forgotten, thereby threatening our ability to present and preserve them. Encouraging the practice of using a standard format vocabulary as suggested, by Abrams and Seaman, for example, should help.²⁶

A further metadata enrichment challenge has to do with supplying information that will assist in the discovery of objects stored in the archive. This process is complicated at the best of times, and is especially complex for large-scale operations in which the volume of objects far outstrip capacity for the development of hand-crafted, object-level metadata records. Accordingly, any large-scale web archiving project will require automated tools for metadata extraction (recognizing structured metadata within captured objects or using heuristics to copy potential metadata such as titles and authors directly out of texts) and metadata generation (using machine learning algorithms to synthesize classifiers and subject headings).

²⁶ Abrams, Stephen and David Seaman. Towards a Global Digital Format Registry. *World Library and Information Congress: 69th IFLA General Conference and Council*. Berlin, August 1-9, 2003. http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf

As indicated above, such tools are imperfectly developed, forcing many projects to adopt a hybrid approach that combines manual and automated metadata assignment. And more complicating still is the fact that different archival collections will be developed with different users and uses in view, and different object-level metadata will need to be provided in support of these varied aims. What convergence we see in the technologies used to identify and record attributes of web-based information objects is unlikely to be reflected in the descriptive data that archives supply to enable resource location, discovery, and use.

A further challenge stems from the fact that the best organization for an archive of web-based objects is not necessarily the one that reflects the objects' proximity to each other at a fixed point in time. Some of the projects we reviewed organized their archives to mirror their appearance and location on the web at the time of their capture. The Internet Archive's Wayback machine is perhaps the best-known example of this. However, some projects re-present content with standard views arranged along subject, title, or author (corporate or personal) lines. And there is interest in mechanisms that enable the persistent re-organization of web-based archives as a means of supporting very different user communities. Re-organizing web archives is not a trivial task. Fully robust, automated approaches are likely to rely in some combination on the fruits of current research into computational linguistics, concept mapping, and the generation of dynamic ontologies. These approaches, though different, automatically review archived materials and in the course of that review generate classifiers assigning them to individual objects. They then feed the classifiers into an inverted index hierarchy (the more sophisticated, the deeper the hierarchy) that can be browsed as an alternative to the chronologically layered, static link structures that are inherited from the capture of original web sites. By-hand methods of classification undoubtedly achieve more flexibility and possibly even greater accuracy (and currently support those web-based archives that do not present material as chronologically arranged snapshots of the web), but are highly constrained by their cost and their inability to scale. As such, they are only appropriate for web archiving initiatives that are and are likely to continue on a small scale.

3.3 Preservation challenges

Preservation appeared to be the most poorly understood activity of those surveyed. Not surprisingly, it engendered the fewest responses in our interviews with project managers. Aside from the relative lack of experience with preservation, only one other common thread emerged through our work. Every one of the projects we reviewed pinned its long-term preservation plans on some broader institutional development efforts. That is, none of the projects are developing a digital archival repository for web-based government information in its own right. Instead, they are working together with local technology units to help design and develop digital archival capacity that works for a broader range of digital information than that represented by the web-archiving initiative.

Given the paucity of experience in the projects, the following comments on preservation challenges draw heavily from our review of existing literature. Since preservation can mean different things to different people, we offer these working definitions as a means of framing the following discussion. By the phrase digital preservation, we mean the set of strategic activities designed to safeguard into the indefinite future the cultural (artistic, scholarly, scientific) heritage

contained in a given body of electronic information. Safeguarding electronic information includes maintaining the viability, renderability, and understandability of digital objects.²⁷ Here, viability is the extent to which an object's bit streams are intact and readable from the digital media upon which they are recorded. Renderability is the extent to which those bit streams can be translated into forms that are usable by humans or processable by computers. Finally, understandability is the extent to which a rendered object can be interpreted and understood by its users. We see five major challenge areas: 1) persistent identification; 2) migration and emulation; 3) size, complexity, and volatility; 4) storage; and 5) hidden data.

Persistent naming

Persistent naming is fundamental to any digital preservation program. Projects have mostly adopted a wait-and-see attitude as standards mature. Meanwhile, because it is not uncommon to see web archives organized more or less as local mirrors that are "isomorphic" to the source web site, naming often reflects some combination of the timestamp of the crawl, plus the original URL. The leading example of this is the Internet Archive's Wayback Machine.

Canonicalization is a challenge with this interim approach. Almost every URL that forms the basis of such a name will have synonyms that a user may attempt to use against the archive without success, simply because the synonym is not byte-for-byte identical to the originating URL. For the simplest name-based archival access, this kind of object addressing is wanting; of course, attribute-based access (e.g., full-text search) is on the wish list for most repositories. Alternate approaches exist. For example, the Nordic Web Archive (NWA), which includes the national libraries of Finland, Sweden, and Norway, harvests materials and assigns identifiers that conform to the URN naming scheme.²⁸ For this purpose, the NWA generates identifiers in the NBN (National Bibliography Number) namespace.²⁹

The traditional debate about whether or not to include recognizable semantics in identifiers also applies to web archives. In the absence of reliable metadata for the end user, there is an understandably strong desire to include some object-descriptive semantics in the identifier. Unfortunately, these semantics don't age or travel well, hence the pressure to push semantics into an expressive and mutable medium, such as a metadata record. We found little discussion in a web archiving context on the question of semantically opaque identifiers.

Some naming problems can be deferred simply because many archives are organized by default to be local mirrors that merely inherit names determined by the source web site. Capture tools (e.g., HTTrack, wget) typically download file and directory structures in precisely this way. This sort of default naming is inappropriate, however, when content in the archive is supplemented periodically (for example, by capturing data from web sites that have already been visited) or when it undergoes any sort of reorganization.

Periodically capturing web-based materials that may already exist in the archive plays mischief with the default naming applied to materials originally captured in several ways. First, if an

²⁷ See the OCLC/RLG Working Group on Preservation Metadata, PREMIS.

<http://www.oclc.org/research/projects/pmwg/>

²⁸ Moats, Ryan, *URN Syntax*. RFC-2141, May 1997. <http://ds.internic.net/rfc/rfc2141.txt>

²⁹ Hakala, Jula. *Using National Bibliography Numbers as Uniform Resource Names*. RFC-3188, October 2001. <http://www.ietf.org/rfc/rfc3188.txt>

object's URL on the source site changes (as happens during a site reorganization) it is hard to detect and correlate with the originally crawled object URL. In the event that it could be correlated (e.g., in order to keep track of a set of object changes over time), isomorphism effectively ceases between the archive and source site, and a new name must be generated for the object. Second, an umbrella object must be created and named that covers all versions of the object captured over time as the site is revisited. Even at the simplest level of maintaining different versions of one object, the default names must be augmented at least with a timestamp indicating when objects on the site were captured.

Things become more complex if the archive itself is subject to reorganization (e.g., object renaming, splitting, or merging) as a result of the curatorial processes. In these cases, new object names will have to be generated per local decisions. Name creation sweeps in a host of traditional naming problems (such as object granularity and object similarity) that have received scant attention in a web archiving context. Granularity refers the level of finer and finer object subdivision beyond which sub-objects are no longer assigned separate identifiers. Similarity refers to the level of increasing dissimilarity, beyond which two similar objects (e.g., one text encoded in Latin-1 and the same text in UTF-8) warrant separate identifiers.

Migration and emulation

Because digital preservation is a relatively new field and web archiving is even newer, there is little guidance on how to choose a preservation strategy. The literature identifies a significant dichotomy between the strategies of migration (converting data to work with changing platforms) and emulation (developing software that emulates the originally intended user experience, and/or converting older software so that it continues to work with changing platforms). While much discussion has occurred on the merits of both, there is not much experience with either. Until this situation changes, both will be candidate strategies for the persistent management of web archives.

Size, complexity, and volatility

In a web archiving context, a whole set of challenges arise from the size, complexity, and volatility of the web. The sheer size of the web implies an enormous amount of data capture just to archive a fraction of it. Even modest web archiving efforts therefore rely heavily on automation, and it is unknown how much human oversight will ever be applied to the captured material. There is a great deal of pressure to improve software tools in an attempt to compensate for the relatively small amount of human review that may be possible.

In an automated context, this reliance on automation enters the preservation picture at the basic level of inventory. Extra effort may be required just to confirm receipt of an object and to make sure that it was recorded faithfully. Experience shows that the HTTP headers describing a download are not always correct. Nor is the return code from an apparently successful disk write always correct; many financial systems, for example, perform checksums on disk blocks to check that the block written was the same as the block submitted. For the purpose of inventory management, the calculation and storage of bit stream digests (e.g., MD5 checksums) along with archived objects is likely to become a requirement of trusted archives.

A base requirement of ongoing renderability is that a captured object be in a coherent, correct format that is renderable from the outset. Formats must therefore be determined and validated at the moment of download; due to volume, this will involve automated means. Tools such as JSTOR/Harvard Object Validation Environment (JHOVE) may come in handy.³⁰ Two kinds of volatility affect web archiving. The first kind of volatility has to do with the highly unstable nature of web sites. This instability can wreak havoc when an archive visits a site periodically, as described earlier under persistent naming. When a web site is reorganized between visits, it is very difficult to track what content has changed because the names (the URLs in this case) that are supplied for the content will have changed. If changing objects maintain the same names between visits, there is still the preservation problem of how often to save changed versions and how to represent them—save the differences ("deltas") or save entire changed objects.

The second kind of volatility has to do with the medium-term turnover in technical standards and practices on the web in general. This is about both frequency of change and diversity. Evolution of data formats (the subject of migration discussions) and of hardware/software platforms (the subject of both migration and emulation discussions) creates problems not only because of rapid change, but also because not every archiving organization has experience with the many different captured formats. The flux may be so great that archives built by automated processes may end up with materials whose formats will never be known in time to collect the information needed to make them intelligible to future generations.

Simple diversity of platform and plug-in requirements for web materials presents a preservation challenge. It is hard enough to automatically detect in web pages the information that recommends or requires that the page be viewed with a particular piece of software (e.g., "this page looks best with browser X"). Accommodating the full range of possible technology requirements may be impossible.

Storage

Digital storage technology is changing rapidly. At the moment, online magnetic disks are cheap and getting rapidly cheaper. At some point it may level off, and the demand for more storage may suddenly jump. If the growth of web archives parallels the growth of the web in general, collections will eventually have to meet storage capacity demands that may be very hard to predict and budget for. Building collections in ways that can be grown incrementally in a distributed fashion so as to take advantage of the storage and other resources available at numerous sites (e.g., using data grid technology) is quite appealing.

Hidden data

Since the beginning of the web, data formats other than plain text have had the capacity to carry hidden data. This is data that is either rendered in a barely detectable way or that is not rendered at all. Examples are legion: "hidden fields" in web forms, unactivated Javascript fragments, unaccessed HTML attributes, unrecognized HTML tags, and URLs of inline images. If the bit stream of the original format is preserved, so is the hidden data. On the other hand, the utility of much of the hidden data may not always be apparent, and there may be cases in which programs

³⁰ See JHOVE (JSTOR/Harvard Object Validation Environment) <<http://hul.harvard.edu/jhove/>> for information about the project to develop an extensible framework for format validation.

fail to preserve information whose value is recognized too late.

3.4 Institutional readiness

Many of the web archiving initiatives we reviewed talked about the challenges they confronted for finding programmatic institutional support for their work with web-based materials. In particular, they cited the difficulties involved in encouraging institutional investment in the deep technical infrastructure that such work requires over the longer term. Already, these initiatives had grasped onto the related facts that (a) such infrastructure was well beyond their project's budget, scope, and technical and financial capacities; and (b) that they were unlikely to attract institutional investment unless the institution could play a role in supplying digital preservation services for more than just web-based government materials. In this regard, it is hardly surprising to find the Australian PANDORA project leading with respect to the build out of its repository capacity. Located as it is within a national library that sees the preservation of digital objects in general as part of its mission, the PANDORA project successfully leverages the institution's investment in general digital preservation infrastructure.

Elsewhere, projects were making some progress towards a level of institutionalization by becoming part of broader digital library programs. This pathway from skunk-works or an experimental project run on the bases of individual enthusiasm (often combined with external grant funding) to a program run systematically as part of an institution's core services is familiar in digital libraries. Indeed, it seems to characterize those digital library initiatives that have successfully sustained themselves over a number of years.³¹

CyberCemetery, the UCLA Campaign Literature Archive, and the Minerva sub-collections are all becoming part of broader digital library programs and benefiting from core institutional investment in staff and technology infrastructure. To what extent institutional commitment will be sufficient to support the construction of the robust preservation infrastructure that web archiving initiatives will require remains to be seen. Project managers are concerned about support, expressed interest in cross-institutional efforts. Repeatedly, project managers mentioned the need for a greater exchange of information about the identification of good and bad web archiving practices, and the collaborative development of commonly required tools and services that might support efforts for crawling, automated metadata generation, organization of archival collections, large-scale content management, and registries of web archives and their holdings. Co-development of crawling tools was a particular concern. As is evident in the final report for the Pilot Project "netarkivet.dk," the needs for development and maintenance in this area are significant:

Continuous technical and professional supervision is required in order to carry out harvesting. The ongoing follow-up will concern both the quality of the material archived and the identification of new formats and functionalities. Even if ready-made technical equipment is used, it is likely to require constant adjustment...

³¹ Greenstein, Daniel and Suzanne E. Thorin. *The Digital Library: A Biography*. Washington DC: Council on Library and Information Resources, second edition December 2002, first edition September 2002. <http://www.clir.org/pubs/abstract/pub109abst.html>

...Professional expertise is needed to identify the location of relevant sites. Detailed follow-up is required, partly to ensure that these relevant sites are included in the collection of star URLs, and partly to ensure that they are actually harvested. Moreover, technical knowledge of harvesting software is required to make sure that technical problems do not disrupt the harvesting process, just as technical expertise is needed for software development, and conservation/curating and preservation knowledge is required to ensure that the right means are used for long-term storage...³²

Our informants from libraries and other memory organizations echoed similar concerns and expressed even graver doubts about their capacities for developing the technical infrastructure required for web archiving on almost any scale. However, they expressed strong interest in continuing their historic roles of building curated collections by extending their collecting purview to web-based materials. The disappearance of web-based content is a matter of grave concern to these intermediaries. Their users need access to today's web-based information now and in the future, and currently there are no reliable means of meeting their needs. Nor are these intermediaries sanguine about the prospects that third party archives will assemble collections that will serve their users' particular interests. In fact, the tendency apparent in web-archiving projects to shape preservation aims and strategies around a series of local and institutional imperatives was evident amongst the intermediaries. At a one-day workshop, professionals from memory organizations with collecting responsibilities in areas with the most significant stake in government information were asked to consider seven models for capturing, curating, and preserving web-based government materials. The models are briefly described in Table 5 below. Each of the models mirrors what professionals in memory organizations do in the print world and none are mutually exclusive.

Table 5: Seven models for capturing, curating, and preserving web-based government materials

Model	Description
Model 1 Click, print, bind	Locate digital content; capture it, print it, and then handle it using existing processes for print publications.
Model 2 Click, save to disk	Locate digital content; capture it, save it to disk. Once saved, decisions about storage, access, and preservation must be made.
Model 3: Targeted crawl (automated)	Capture content according to particular criteria—for example, all files modified after a certain date, with particular metadata, in a particular format—and save to disk.
Model 4: Interactive crawl (facilitated)	Initiate a crawl and save to disk. As the crawl proceeds, reports are issued that require input. The crawler is prompted to accommodate this input. This process is repeated as the crawl proceeds. As the crawler encounters a problem, an event log is created for human review. For example, the crawler encounters a form, prompts a human for help, the human fills in the form, and then the crawler captures the site.
Model 5: Negotiate with producer	Negotiate with the data producer to acquire content and save it to disk (potentially web and non-web content).
Model 6: Save blindly	Perform global capture, without prior selection or analysis activities other than targeting a domain or set of sites, and save to disk.
Model 7: Save blindly, then select	Perform global capture and save to disk. Select materials on the stored content, ex post facto.

³²Christensen-Dalsgaard, Birte et al. *Final Report for The Pilot Project "netarkivet.dk"*. 2003. <http://www.netarkivet.dk/rap/index-en.htm>; pp 59-60.

Participants were asked three sets of questions related to each model: what are the advantages; what are the disadvantages; and what are the key functional requirements (presentation, access, cataloging and preservation) if this model were implemented. In answering these questions, participants were asked to consider various needs, notably:

- ❑ needs of the selectors responsible for building collections of government information or augmenting them with other materials;
- ❑ needs of public service staff responsible for encouraging and supporting use of government collections;
- ❑ and needs of senior managers responsible for strategic decisions about institutional mission and resource allocation.

In the ensuing discussion, a number of things became clear. First, participants' preferences for particular collection development models were determined by the communities they served, their institution's historic collecting strengths and collection-building missions, and by the resources (technical and financial) that were available to them. Not surprisingly, no collection development model was preferred over others by a majority, or even a plurality of participants. Second, the animated discussion around model 7, which involved an immediate and far-reaching crawl of the dot-gov web domain in advance of any clarity about possible curatorial and preservation tactics to manage the captured content, indicated the sense of urgency the group felt about the at-risk nature of government information. Participants that favored this model saw it as an important safety-net for information that was deemed to be as volatile as it was valuable. Model 2 appealed to participants because it empowered individual curators or selectors to act in advance of more comprehensive preservation strategies and technologies being put into place. In effect, it enabled them to build personal collections of web-based information that were deemed by the selectors to be the most important. Model 4 was also attractive because it occupied a middle ground between the global and unmediated capture strategy of model 2 and the personalized book marking approach of model 7. A full list of the results of the discussion—the advantages and disadvantages of each model, and suggested functional requirements—is available as *Appendix 3*.

A great deal of urgency has been expressed by librarians and other intermediaries with regard to the rescue of web-based government materials. Along with this urgency is considerable frustration at the absence of tools and services that enable them to fulfill their curatorial roles without their institutions having to build the technical infrastructure and harness the expertise typically associated with data warehouses and data archives. In this regard, the librarians and other intermediaries we interviewed, like the project managers involved in web archiving initiatives, expressed enthusiasm for partnerships and other collaborative arrangements that would allow them to combine their curatorial strengths with the technical capacities available at other institutions.

Conclusion

Web-preservation practices vary considerably from one archiving institution to another. The decisions that govern what web-based information is selected and captured for inclusion in an archive, and how archived information is described, is influenced by a variety of local issues about the archiving institution's mission, its existing collection strengths, and the needs of its

users. This variation seems desirable. No institution is able or willing to capture the entire government domain. Even the Internet Archive is selective; it misses the deep or hidden web and implements some format constraints. Redundant archiving practices promise to extend the breadth of web-based materials that are brought into persistently managed collections.

Redundancy is also valuable since it promises to meet a broader range of user requirements. As we have seen, different archiving initiatives will define the government domain differently. They also enrich collections with different metadata schemes and in doing so, promise to support particular users or uses. Redundancy of practice is also valuable where two archives capture the same content and manage it differently. If the preservation strategy implemented by one fails to stand the test of time, materials in its collection that are managed differently elsewhere may yet have a chance of survival. Redundancy seems particularly desirable given the fact that web archiving is a new and challenging activity about which a great deal remains unknown. Variation in practice is an essential means of surfacing viable best practices.

There remain, areas where redundancy is clearly inefficient, and these were often front-most in the minds of our informants. Virtually all of the preservation initiatives we surveyed were struggling to supply themselves with the same tools, such as web crawlers (particularly where these offered a high degree of local configurability) and automated metadata generation tools. Few lacked the research capacity necessary to even assess promising new technologies, let alone develop their own. Informants frequently expressed their desire for venues where they could share information with other web-archiving initiatives, and consortial efforts through which they could pool scarce research and development efforts. Additionally, informants at most of the projects we surveyed expressed their very serious concerns about the prospects of having to build a data repository that would be capable of persistently managing their collections, which they expected to grow exponentially in size. Most sought to piggyback their requirements on enterprise-wide initiatives undertaken to persistently manage a broad range of digital information, but here too there were doubts about the level of commitment and capacity.

These findings challenge us to think about mechanisms that:

- ❑ sustainably support a wide variety of preservation efforts at institutions that combine historic experience safeguarding the government record with an expertise in the development of collections that satisfy the needs of particular user communities;
- ❑ lower both the cost and the risk to those institutional efforts by supplying them with the common range of tools and services that they require but cannot independently afford.

The service model in the next section describes those mechanisms.

4 Archiving the web: a “layered” service model

The nation’s government publications are clearly at risk. There are many reasons, not least the scale of the problem, which is perceived as so vast as to defy meaningful contributions by the libraries, government agencies, and other memory organizations that are best suited to contribute

meaningful solutions.³³ This perception has to do, in part, with how digital preservation is modeled by those memory organizations: as a complex process that needs to be managed end-to-end within one archive. The model works well for the paper-based and analog archives from which it derives. There, the preservation process is organizationally centralized by necessity in archives of physical objects.

This centralized model is not essential to the persistent management of digital information because the physical location of that information is inconsequential. Digital collections can be built and described by one organization, managed by another, and delivered to end users by still another. This kind of organizational division of labor by function is not only possible with digital information, it is desirable because it leverages specialist capacities that tend to be organizationally distributed. It is also not uncommon. Most online content services are not responsible for the assembly, management, and delivery of the digital information that they offer to clients as their own. For a library, such offerings extend to electronic journals, databases, and e-books.

Still, in their thinking about archiving, libraries, government agencies and others that have a stake in the persistent management of web-based government materials remain wedded to this outmoded archival service model. So long as that remains the case, the digital archives established to maintain our governments' web-based outputs will remain few and far between, cropping up only at those institutions that are able to supplement their curatorial activities with the vast technical infrastructure that is essential to a digital archive. Given the impacts of scale, such initiatives are also likely to be very highly focused, dealing with particular, but relatively small corners of the government information domain much like the projects that we surveyed in the course of this research. They may in fact be restricted:

- ❑ to commercial, not-for-profit, and government agencies that are mission-driven by legal, fiduciary, business planning, or other needs, to persistently manage the digital assets they create (e.g., the U.S. National Archives and Records Administration, the U.S.'s National Archives and Records Administration, and selected publishing and entertainment companies); and
- ❑ to that handful of academic, research, and public libraries that can viably justify a claim as guardian over some small segment of the government record (as is the case in the National Library of Australia, University of North Texas' CyberCemetery, and the Library of Congress's Minerva project).

While we would not wish to discourage the development of new, institutionally based silos of activity, real progress may require that we find creative ways to align organizationally distributed technical capacity with the wealth of curatorial expertise and archiving enthusiasm that is widespread in memory organizations. Here, the information architecture proposed by the Library of Congress for its National Digital Information Infrastructure Preservation Program may provide essential guidance.³⁴ It defines the preservation process as comprising a number of

³³ Lavoie, Brian F. (2003) *The Incentives to Preserve Digital Materials: Roles, Scenarios, and Economic Decision-Making*. White paper published electronically by OCLC Research. Available online at: <http://www.oclc.org/research/projects/digipres/incentives-dp.pdf>

³⁴ Clay Shirky, Appendix 9. Preliminary Architecture Proposed for Long-Term Digital Preservation. In Library of Congress, *Preserving Our Digital Heritage. Plan for the National Digital Information Infrastructure Preservation*
Web-based government information: capture, curation, preservation -- California Digital Library

discrete roles, including those of *repositories* (store information), *gateway* (manage the flow of information into and out of repositories), *collections* (selectively assemble digital information and determine its characteristics and the terms of its use), and *interfaces* (build end user and other services on top of collections, offering a modicum of access to them). These roles can be co-located in digital archiving silos (as they presently exist in a small number of organizations). Or they can be distributed in a way that layers one service on top of others. Our principal findings in this research suggest that the organizationally distributed and layered model is preferred. In particular, we reference the immense variability of memory organizations' archiving aims and strategies for web-based government information, and the impediments those organizations face in the absence of ready access to common archiving infrastructure, services, and tools.

In this section, we provide a broad outline of this layered service model and how it might empower memory organizations to take up their historic roles as guardians of our governments' published outputs by lowering the costs involved for acquiring access to commonly required tools and services. In a following section, we outline a preliminary and much more detailed analysis of what key service components might consist of.

4.1 Mapping roles and incentives to organizations

Repositories and agents

Digital preservation ultimately requires professionally managed large-scale storage facilities and the means of managing the flow of data into and out from them. The costs involved in developing and managing such "bit farms" are significant and are creating real barriers to the memory organizations that are best suited to the capture, curation, and persistent management of web-based government information. Of the initiatives surveyed, only the project at the National Library of Australia could lay any claim to the technical infrastructure that it required. Elsewhere, projects pinned their hopes on large-scale investments in general-purpose repositories through which an institution could manage a broad range of the digital assets, including the web-based government materials that they were actively capturing. Although there is evidence of some movement in this direction—three of the projects we surveyed had been or were in the process of being subsumed into larger digital library initiatives—the final result is clear and is, indeed, the source of considerable concern among the project managers.

While library and other memory organizations face real obstacles in surfacing the deep technical infrastructure they require to support their effective preservation of web-based government materials, institutions that support high-end, large-scale computing facilities appear very reluctant to take on the preservation challenge. They are able to spin bits but unlikely to undertake the curatorial challenges involved in determining what bits to spin and how those bits should be described, organized into collections, and later rendered intelligibly to meet the needs of different end-users. This perspective, at least, emerged very powerfully in our work with the San Diego Supercomputer Center (SDSC). Colleagues at the SDSC enthusiastically supported our use of their data storage and grid technologies, but insisted that the CDL team direct its

Program. A Collaborative Initiative of the Library of Congress, October, 2002).
http://www.digitalpreservation.gov/ndiipp/rep/ndiipp_appendix.pdf

application. In other words, staff at the Center were interested in contributing technologies to digital preservation efforts where key selection and curation decisions had already been made.

The service model proposed here reflects, to a large extent, our experience working with the SDSC on the current project. By seeking some organizational distribution of preservation functions, it may be possible to leverage the substantial bit farming capacity that already exists at SDSC, at other national super-computer centers (Illinois, Colorado), at universities with substantial IT infrastructure (Indiana), and at commercial entities that manage large-scale computing facilities to support their research, product development, or service missions (e.g., in the telecommunications, Internet service, and defense contracting industries). The bit farmers could, in this scenario, act as SDSC did in our very limited pilot project, that is, by building at marginal additional cost on what they do best without incurring the additional responsibilities (and costs) associated with designing, curating, and delivering archival collections to specific end-user communities. Incentives to the bit farmers for making additional investment are clear enough; they create additional demand (and potentially provide additional revenue for) the technical infrastructure that is largely in place.

Collections and collectors

Persistent collections of web-based government information are built to very different designs as determined by a complex array of local mission critical issues that ultimately determine what information is captured, how it is described and organized, and how, to whom, and under what circumstances it is made available. And within memory organizations that have historically safeguarded the published outputs of U.S. governments there is no shortage of enthusiasm to transition their role into a digital realm. Yet their aspirations are stymied by the costs involved in developing and maintaining the necessary supporting infrastructure. Reduce the cost, the argument runs, and more collectors will extend their attention from the governments' printed materials to their web-based outputs. Their incentives for moving in this direction are the same as those that have driven their work with print materials: service to a specific user community, historic concentration, and collection strength in a particular area.

The preservation costs that can be reduced for collectors were clearly highlighted in our interviews with both project managers and with intermediaries, and include the costs associated with bit farming. As already indicated, large-scale technical infrastructure is not easily or readily replicated. As such, it represents a substantial barrier to the development of persistent collections. Other costs that keep collectors out of the preservation business and that can be substantially reduced (or entirely eliminated) include those that are involved in:

- ❑ developing, maintaining, and becoming familiar with the use of essential tools that enable the curator to identify, select, capture, index, and migrate the digital information they wish to archive, and to do all this in a way that meets local needs and interests;
- ❑ establishing, licensing, and managing relationships with bit farms;
- ❑ ensuring that collections content continues to integrate with the rapidly evolving software applications necessary for accessing or using collections content;
- ❑ keeping in touch with relevant applied research and development efforts; and
- ❑ providing local services that really ought to be supplied commonly to anyone acting in a curatorial role (e.g., name resolution services that ensure all digital files are uniquely identified and identifiable and auditing services that build credibility around persistent

collections), and to registries (that indicate who has what, and by doing so, ensure that content redundancy is planned and needed rather than serendipitous and wasteful).

The roster of potential collectors is far longer than that of bit farmers. It has to be. To effectively preserve web-based government information, we must encourage redundancy in practice as well as in content. Content duplication is known to be good; it builds a failsafe mechanism to protect against the catastrophic loss of a single archive. In addition, we want the digital archives to manage their redundant holdings in different ways, that is, to use different data capture techniques, archival formats and format migration or emulation schemes, different levels of data description, etc. Digital archiving practice is tried but not thoroughly tested. Diversity of practice builds a failsafe mechanism in case a practice proves ineffective or unsustainable. It is virtually assured if the libraries and other memory organizations that have an interest in the persistence of web-based government materials are empowered to build archives to their own scope and design.

Reducing collectors' costs sufficiently to encourage widespread participation in a highly distributed network of collection efforts is possibly the largest obstacle impeding the persistent management of web-based materials.

Interfaces and user services

Although access to archives of web-based materials is beyond the scope of this report, it needs to be accounted for in the service model. The electronic information industry provides adequate evidence that the mere availability of digital information is incentive enough to organizations that build and manage end-user and other access services. Whether producers of digital information distribute their content openly or under license, freely or for money, the majority wants as many pathways as possible leading to their work.

Enabling the development of numerous end-user and access services (above and beyond those that the collector will develop) must be a primary aim for collectors. The tendency to make digital collections available for development by third parties into "higher-level" end-user and access services is already apparent in the information industry, for example, at Amazon.com and Google, among commercial journal publishers through their reference linking services, and among digital libraries through their virtual union catalogs. It is also apparent among existing digital archives, many of which are frankly relieved of having to simultaneously service countless and varied user-service demands while acquiring and persistently managing collections.

Trusted preservation brokers

The role of the trusted broker is not addressed specifically in the Library of Congress's information architecture. It is an organizational rather than a technical role for which the need became increasingly clear in the course of our investigation. The broker's role as we envisage it is to stimulate the economy that will sustain the interdependent efforts of repositories, agents, collections, and even interfaces. Brokers may:

- ❑ enlist collectors focusing on different aspects of the government domain or focusing in different ways on the same content;
- ❑ put collectors and bit farmers in touch and assist them in defining their relationships;

- ❑ supply collectors and bit farmers with tools and services that lower costs, improve effectiveness, and ensure redundancy of collectors' practice (tools for crawling the web and for automatically assisting metadata generation were particularly sought after by our informants);
- ❑ surface and articulate demand (e.g., from collectors and bit farmers) for new tools and services (persistent naming utilities and registries of archives and their holdings emerged from our interviews), and pool investment in their development; and
- ❑ arrange and ensure appropriate accreditation and validation of preservation efforts.

Implementing the preservation broker's role

The following scenario illustrates the potential value that a trusted broker can contribute to the preservation process. An organization offers a range of networked library services. It has its own technical infrastructure, including a small group that focuses on applied research with cutting edge technologies. It also maintains a network of contractual and strategic partnerships with:

- ❑ content owners and distributors;
- ❑ large-scale computer facilities and data services;
- ❑ computer science research facilities; and
- ❑ the broader national and international standards and professional communities that have grown up around digital libraries, digital preservation, information interchange standards, and related efforts.

Driven in part by its clients' needs to capture web-based materials and extend their historic government collections, the broker seeks to install the tools and services that will enable the client to do so in a cost effective manner. Working among its clients, the broker develops a functional specification that suggests its clients are interested in preserving different but overlapping sets of networked information. It also reveals their interest in employing different preservation strategies and practices, even where they focus on the same digital content. The exercise also suggests that clients will take on the collector's role if they have low-cost access to secure repositories and to a suite of tools and other support services that enable them with a limited amount of local technical capacity to analyze, capture, organize, index, and deal fairly with the copyrights in a manner that supports their preferred preservation strategies.

In light of this information, the broker builds upon existing relationships with data centers able to act as repositories. It also leverages its contacts in the information industry to identify and acquire a set of tools and the implementation guidelines that will support clients in their use. For those tools that are unavailable (either in industry or the research sector), the broker works through its research associates (in industry but also in academia) to develop them.

Networked information resources are represented in the upper right corner as a morass of information content with different characteristics. The broker then supplies, or negotiates the clients' access to a suite of tools that enable the client to build persistent collections that meet their needs. Those tool suites are discussed briefly below:

- ❑ *Analytical suite* enables client collectors to analyze the web-based government materials they are interested in (e.g., to determine scope, depth, and characteristic), and use this data to determine optimal data capture, curation, and preservation strategies.
- ❑ *Data capture suite* enables clients to acquire digital information for their archives in a way that serves their local needs. A variety of tools would be available, each with a

degree of configurability. The tools support highly granular selection (such as single files, single categories of files, or MIME types) and more broad-based approaches (such as large web sites or sets of web sites within a particular range of domains). Another suite of tools supports curators who receive digital information directly from the producer, owner, or distributor.

- ❑ *Preservation and storage suite* is required by curators to store captured digital information. They include a holding location for assessing material during and immediately following capture. Depending on client requirements, the broker might offer the delivery of archival files to the client for inclusion in its local repositories (whether that is a local file server or institutional repository), storage in a shared public space managed by the broker, preservation in the broker's own repository, or transfer to another utility services. A client might desire to combine one or more of these approaches.
- ❑ *Curatorial suite* helps curators index and organize their collections, ensuring that content could be located and made readily accessible in accordance with any access restrictions. The services range from cataloging and classification to an array of automated metadata generation tools. Related options for access to the selected content will be provided to subscribers to this suite.
- ❑ *Administrative suite* provides curators with the information and tools needed to construct the policies, practices, and rights frameworks that will govern their selection, curation, and preservation activities. For example, the suite includes the provision of guidelines and standards for the management of collections of web-based content, sample licenses for collections requiring capture mediation, sample data deposit and data transfer schedules, and model shared intellectual property agreements, among other things.

A further key element is a registry of web content already captured and preserved by various clients that the broker serves.

The broker also plays other roles: for example, liaising between its own network of clients, repositories, end users, and broader digital preservation research and development communities. The broker also represents its network in defining specifications and seeking sustainable support for essential national and international utility services (also not represented) such as name resolution, registry, and auditing services.

The scenario demonstrates the value of the distributed organizational model, which is able to achieve the following:

- ❑ It enables organizations that have a historic and mission-critical interest in developing persistent collections of digital assets to develop those collections, notably by lowering both the costs and the risks involved.
- ❑ By building a supply of collectors, it creates demand for repository and agent services and leverages existing capacity to supply those services while alleviating the responsibility of also having to fulfill curatorial roles.
- ❑ It offers theory- and content-neutral support for those developing digital collections. In this regard, it empowers collectors without dictating how they should build persistent collections. This does two things: it builds redundancy in archived content and in archival practice, and it provides incentives to collectors who are able to build collections that meet their local user and market interests.

- It provides an extensible basis of operation. The network of repositories and collectors developed around the broker can extend indefinitely (for example, to include collectors from other libraries, but also from the commercial and other sectors) as can the tools the broker helps to define and supply.

Brokers are common in industries that seek to overcome imperfect markets or that tend toward vertical integration. They are familiar in the information industry (where publishers, for example, broker the intricate relationships between authors, editors, printers, and wholesale and retail distributors that ultimately contribute to the book trade) and in the heritage sector (where consortial, membership-based, and other non-profit entities ensure that libraries and museums have access to a range of services that they require but cannot independently afford).

The incentives for brokers to emerge in the roles described above are likely to be varied. Many will be organizations that already serve collectors, repositories, or research and development efforts in some way and for which the brokering role aligns closely with their current missions. Examples might include library utility services (OCLC, Research LG, CDL, OhioLink or even Los Alamos National Lab), government agencies that have some legal or other obligation to facilitate persistent access to web-based publications (such as GPO and state-level equivalents and state libraries). Some might emerge as new organizational entities, as did JSTOR, a scholarly journal archive, in an earlier era to pool investment in the digitization and preservation of selected scholarly journals.

5 A route map for service implementation

When fully matured, institutions participating in the various roles identified in this layered model should sustain one another economically. The existence of bit farms (repositories and agents) will encourage curatorial institutions to take up their historic preservation roles with digital materials. As curators proliferate, they will create demand for repository and agent services sufficient to sustain them. As archival collections proliferate in number and grow in size, opportunities will emerge to create new and highly innovative end-user services, which may contribute tangibly to the support of underlying curatorial and repository efforts.

The challenge then, is not in sustaining the economy surrounding interdependent organizations that collectively contribute to preservation as an organizationally distributed process; the challenge is kick-starting that economy. Here, the broker's role may be particularly important. As envisaged, it has capacity to align the currently independent efforts of curators and repositories and to lower or at least make more predictable the costs involved to each in undertaking new and inherently risky initiatives.

The following section charts out initial steps that CDL will take to kick-start the service economy, notably by developing itself in the role of broker as outlined above. Although the discussion is highly specific to the CDL, it is presented in terms that are general enough to provide a route map for others that may wish to follow this path.

In its brokering role, the CDL will need to ensure that client collectors have access to both the repository capacity and the tool suites described above. Accordingly, it will undertake to build, acquire, or arrange access for collectors to these suites in the manner described below. In an initial service implementation, collectors would use the various tools supplied by or through the CDL in order to capture web-based government materials and curate and manage them in a repository maintained by the CDL. In later implementations, it is anticipated that some collectors may prefer to manage the content they capture in their own local repositories or in repositories maintained by third parties. Since the absence of available deep repository infrastructure seemed to be a major impediment to our informants that want to build persistent collections of web-based materials, we feel it is an essential element of any service infrastructure that a broker will need, at least initially, to provide access to.

To capture web-based materials, collectors will use an interface that allows them to download web pages or entire web sites into the repository, subject to certain limits and basic sanity checks. The collectors' web archive will consist of the pages they download. The CDL will also be able to present a federated view of collectors' archives including the union of all of the web-based material that they collectively download.

Curation for the archive will focus on the activities supporting public browsing and searching of the web archive as a collection of mirrored web hierarchies. Simple mirroring reflects the provider's original intention and avoids the cost of manually reorganizing a site's content each time it is crawled. This does not preclude subsequent curation activities, funds permitting, that allow for reorganizing and reformatting crawled content.

Search indexes will be built from page content (full text), from metadata gathered as a by-product of the crawl (such as file types, sizes, and URLs), and from any metadata that's easy to automatically generate or extract. Examples of the latter might include titles and subject classifiers assigned by a curator prior to a crawl that would be attached to the entry URL of the crawl.

Preservation activity will be restricted to periodic bit refreshment and periodic migration of material to new platforms as old platforms cease to be viable. Long-term persistent identifiers will normally be based on the original crawl timestamp and URL of the resource (similar to the Wayback Machine). Where curation of an object calls for creating multiple snapshots, a persistent identifier will be generated to name the composite object. The migration strategy will make the entire archive accessible from active production systems (perhaps with suitable barriers to prevent less relevant material from being included in the default scope of a search), which has the benefit of keeping web archive objects "on the radar screen" of production system migration planners.

The following core service components comprise the web archiving vision and are slated for development in the first service implementation.

- ❑ A repository infrastructure capable of managing both large- and small-scale collections of web-based materials as defined and collected by client libraries
- ❑ An entry Point URL (EPU) Registry

- ❑ A curator Interface
- ❑ Registry Content Search
- ❑ Archive Content Search
- ❑ Crawler
- ❑ Indexer

Each of these service components is described below along with an indication of some of the more promising tools that may be used to realize them in production. Unless otherwise indicated, tools were selected according to the following criteria: 1) an open source code base; 2) scalable to the volume of materials expected from ongoing bulk capture operations; and 3) subject of reasonably favorable reviews. For some tools, the existence of reviews was critical because the number of tool choices was so large it would be difficult to evaluate each one in-house. In some cases (e.g., the registry database), the application area was sufficiently well understood that tool selection became less a matter of research and more a matter of choosing a tool that integrated with existing organizational infrastructure. A complete list of the sources consulted when reviewing and selecting specifics is included in *Appendix 4*.

5.1 Repository infrastructure

The web archiving service relies on a robust and highly reliable repository infrastructure. To create this infrastructure we designed a repository with extremely reduced functionality to make it all the easier to analyze, secure, verify, troubleshoot, and migrate. The design also supports replication and incremental addition of storage capacity. To increase our confidence in the system, the architecture and source code will be open, redistributable, and extensible. One test of design will in fact be the eventual extension to a generalized repository framework.

The repository infrastructure will be built by connecting a number of standards and existing system components. Generally speaking, the interfaces and components rely on concepts from the OAIS (Open Archival Information System) model.³⁵ The repository design also uses METS (the Metadata Encoding and Transmission Standard)³⁶ for creating AIPs and the ARK (Archival Resource Key)³⁷ naming scheme to identify them. Tools for use of METS and ARK are already under development at the CDL, and will be refined and extended for the web archiving project. Finally, Shibboleth (a project of the Internet2 MACE group)³⁸ will likely inform our authentication strategy.

At the lowest level, the design rests on a grid of storage servers at each participating institution. Each localized grid is composed of "bricks", which are low-cost computers with attached storage; the computer and its storage fits in a small, uniform physical chassis that can be

³⁵ The Open Archival Information System (OAIS) Reference Model is currently being reviewed as an ISO Draft International Standard; see <http://ssdoo.gsfc.nasa.gov/nost/isoas/us/overview.html>

³⁶ For information on METS see the official website: <http://www.loc.gov/mets>

³⁷ For an overview of ARKs see Kunze, John A. Towards Electronic Persistence Using ARK Identifiers. *3rd ECDL Workshop on Web Archives Semantic Web*. Trondhiem, Norway, August 21, 2003. <http://bibnum.bnf.fr/ecdl/2003/proceedings.php?f=kunze>

³⁸ for information on Shibboleth see: <http://shibboleth.internet2.edu/>

"stacked" (rack- or shelf-mounted) in the modest-sized machine rooms that tend to be available to participating archiving institutions.

The bricks in a grid are tied together by an overall database server that allows all the bricks in one grid to be viewed as one coherent collection of storage units. This illusion is accomplished by having all access intermediated by a simple database server that maps persistent identifiers to physical storage locations. The database server is essentially a gateway between a collection and the repository.

The primary collection-initiated functions supported by a local grid are extremely simple. The ingest function envelopes an incoming object in a METS wrapper, makes sure it has an ARK identifier (minting an identifier and binding it to the object if need be) and pushes it into storage. The access function accepts an ARK identifier, and returns the object bound to it. The logging function keeps a record of each ingest and access for the purpose of producing management reports. Aside from a simple authentication mechanism and the communication protocols, Phase I requires no more than this.

While each institution may choose its own low-level storage grid provider, at least two of them will use the open-source Storage Resource Broker (SRB)³⁹ system. For those using SRB, it is very easy to add a new brick (e.g., to add a terabyte of disk storage) to a local grid as needed without disrupting current operations. It is also simple to configure an SRB grid such that each deposit of data into the grid causes an automatic deposit of the same data into the grid of a cooperating institution; in this manner, a high degree of geographic replication can be achieved.

5.2 Entry Point URL (EPU) Registry

An Entry Point URL (EPU) Registry is a database table that records the starting point of a collector's crawl and related information about who (name, institution, email) initiated it, a curator-assigned name for the EPU (not necessarily unique), when the crawl was made, the number of pages it returned, at what depth and with what breadth it was made, etc. The EPU may be used to identify one complete document or the root of an entire web site. Hence, the crawl of an EPU may result in the retrieval of anywhere from one page to thousands of pages or objects.

The Registry would serve several purposes.

- ❑ It would help ensure that collectors did not inadvertently or unintentionally capture material redundantly. An attempt to crawl a previously crawled entry point would trigger an exception in the curator interface.
- ❑ It would be used to implement user-based browsing and searching. For example, was this URL saved in the archive?
- ❑ It would be the clearinghouse for all archive activities. As such, it reflects the current inventory of crawled EPUs at any moment in time.

The Registry must also eventually accommodate requests to remove material as copyright protections are discovered.

³⁹ See <http://www.sdsc.edu/DICE/SRB/Pappres/Pappres.html> for recent publications and presentations on SDSC's Storage Resource Broker.

Since the EPU Registry would act as a clearinghouse for all archive activities, the actual record composition would change to accommodate additional service features as they become available. For example, a new data element might be added to the registry to identify how often an EPU is crawled automatically. An important general feature of the registry, therefore, is a flexible and expandable definition of record data elements.

Robust archive operations require procedures and practices already commonplace in libraries and data centers at a minimum. Beyond this, core components such as the Registry and the archived objects themselves would normally be held in transaction-capable databases. Accreditation mechanisms and policy development are also planned to ensure confidence in the registry and the archive it represents.

Candidate tools for EPU Registry

By itself, the Registry is a straightforward application that requires few special-purpose tools. Scalable, open-source tools that integrate with existing organizational infrastructure apply here:

- ❑ Open source MySQL with BerkeleyDB tables <<http://www.mysql.com/>>
A fast, proven, open source relational DBMS with an active development community. The recently incorporated support for BerkeleyDB tables makes MySQL transaction-safe.
- ❑ Storage Resource Broker from the San Diego Supercomputer Center
><http://www.npaci.edu/DICE/SRB/>)
Client-server middleware that accesses heterogeneous replicated data stores through the intermediary of an attribute-based Metadata Catalog (MCAT).

Sample policy documents:

- ❑ Trusted Digital Repositories: Attributes and Responsibilities
<<http://www.rlg.org/longterm/repositories.pdf>>
Articulates a framework for handling the range of materials held by research institutions while providing a basis for the expectations of a repository that is reliable and trustworthy over time.
- ❑ California Digital Library Persistence Policy
<http://www.cdlib.org/programs/digital_preservation.html>
Defines the nature of the CDL's commitment to the digital objects in its care.

5.3 Curator interface

The curator interface to the EPU Registry is a service that remotely located curator and administrative users will need to initiate and monitor their crawling operations. Access to this interface must be controlled at the level of the client (collecting) institution. Ordinary web server accounts (such as userid and password) should be sufficient to provide accountability and allow for the imposition of any resource limits that may be required. Without such limits, one institution's crawl might inadvertently consume most of the central archive's storage. Automated "sanity checks" during crawls would also be required to prevent a crawler from wandering out-of-control with adverse impacts on either the repository or the crawled targets.

Once past the password challenge, the interface would present a simple form for the collector to fill out, asking questions such as what entry point UIRL (EPU) the crawl will start with how deep (to what level) to crawl, what file types to collect, etc. The collector would also supply an email address where crawl results would be sent. This is important a crawl may take a long time to complete thereby making normal web interaction impractical.

Curator-assigned EPU metadata

A possible enhancement to the interface would be to let curators assign metadata at the EPU level before the crawl occurs. The set of metadata would include ordinary descriptive elements such as title, author, and date; especially important in determining the relevance of these would be an element indicating whether the crawled EPU is a work or an aggregate site. For example, it would be easier to assign a subject classifier (e.g., from LCSH or DDC) to a single work or a highly uniform aggregation of works. Mixing these capture and curation operations in one interface is justified because for manual operations, it's more efficient to handle objects once rather than to go back to add metadata later on.

Recording the existence of locally saved crawls

With another possible enhancement, the curator interface could be used to record information about collectors who, armed only with a standard web browser, acted to save a set of web-based materials to non-shared media. The feature would enable the broker service to record information about materials being captured by individuals working on a small, often highly personalized scale, even though the materials would not be saved in the shared repository. The feature would effectively support the variety of localized individual efforts that we uncovered in our interviews with staff in libraries and other memory organizations. The existence of a local copy of some web-based material is of sufficient general interest (e.g., imagine an interlibrary copy request) that the EPU Registry would be a natural place to track its existence and avoid duplicative capture. There are, however, limitations in this service enhancement. Beyond limited discovery mechanisms based on registry records (see the Registry Search component), access to such locally saved resources might be strictly limited.

Federating registries

Another enhancement would support federation of EPU registries where these are maintained by client institutions as a means of monitoring and keeping an inventory of their local web-collection efforts. This would support the development of autonomous repositories that take advantage of local policies and storage resources while ensuring that information about those repositories and their holdings are integrated with like information comprised in the CDL's central EPU registry. To support unintended duplication of archiving effort and some limited discovery by the wider community, local registry record changes would need to be sent to the central registry.

Under this federation option, the central registry would contain two kinds of records: those under control of the curator interface supplied by the CDL, and those (read-only) that it would import from a set of known local registries. This functionality would require implementing a separate communication protocol to import records and augment the authentication regime to define the identities of known local registries.

Candidate Tools for Curator Interface

Like the Registry, by itself the Curator Interface would be a fairly straightforward application that would require few special-purpose tools. Scalable, open-source tools that integrate with existing organizational infrastructure make sense here, such as.

- ❑ Open-source Apache web server
<<http://httpd.apache.org/>>
Reliable, open source web server software. An ordinary configuration would suffice (e.g., no need for strong authentication).
- ❑ Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
<<http://www.openarchives.org/OAI/2.0/guidelines-oai-identifier.htm>>
For the federated Registry case, this would be a possible protocol for importing records, provided record harvesting is set up to be frequent enough (e.g., nightly) that the federation is perceived to be coherent.
- ❑ PANDAS web-based management system from the NLA
<<http://pandora.nla.gov.au/manual/pandas/>>
Allows different participating agencies to: 1) create and maintain records for all titles, collections of titles, publishers, and indexers; 2) search for records using a number of options; and 3) initiate the archiving of individual titles.

5.4 Registry Content Search

The Registry Content Search service component is one of two major components of the web archive user interface. Registry Content Search would be based on indexing data found directly in registry records. It would be complemented by the Archive Content Search, which is based on URLs and on words found in the downloaded objects. Although the user interface should minimize the boundaries between these two components, the Registry Content Search is conceived separately because:

- ❑ its implementation is much more straightforward than the Archive Content Search (the search and browse of the registry is based on the registered EPU or collection level, rather than the crawled object level); and
- ❑ the Registry Content Search provides the natural basis for system administration.

At a minimum, this search component would allow users to search for EPU records by various data elements and in different combinations. It would need to support a "contains" query that would allow users to enter a URL and have the system return EPUs that name an initial substring of the URL. For example, it would need to return the EPU as the parent of a crawled hierarchy that contains a "snapshot" copy of an object that was once at a user-specified URL.

Implementation would require a Btree index and access could be delivered to users through a simple form-based interface.

Simple record display should support most browsing requirements as envisaged for the registry. A small registry with discernible ordering (e.g., alphabetic by EPU) could be browsed simply by paging through a record set that contains all records. If the registry becomes large enough to warrant it, subsets could be pre-defined (e.g., the results of canned searches) and presented as browsing choices.

An administrative variant of this interface (authenticated either by password or by IP filtering) would allow searching on more elements and some administrative options.

Candidate tools for Registry Content Search

Registry Content Search is another straightforward application that would require no special-purpose tools. A generic, scalable, open-source tool that integrates with existing organizational infrastructure makes sense:

- PostgreSQL and MySQL open source database management systems
<<http://www.postgresql.org/>> and <<http://www.mysql.com/>>
Both are fast, proven, open-source databases that will support a simple search of the data in the registry.

5.5 Archive Content Search

Archive Content Search would be the second major components of the user interface. It would be based on indexing URLs, full text words, and other data found in downloaded objects, and would be complemented by the Registry Content Search as described above.

A simple form-based query interface is envisioned. Search results would be a set of matching objects in the archive. A record for each object in the archive would be generated when it was downloaded during a crawl (see the Indexing section). Record display would be based on elements produced as a by-product of the crawl (e.g., URL, file type, size). Optionally, the record may be created with additional elements (e.g., an HTML title), or automatically generated (e.g., an LCSH subject heading).

Browsing archived objects is envisioned as being similar to browsing a mirrored web site where the source URL is identified with an EPU. It remains to be seen whether automatic classifiers could be useful in creating groups of web-based content browsable by keyword. In the meantime, the Registry EPUs provide natural pathways into the hierarchies of web-based materials that are captured by individual web crawls that could then be explored with ordinary client-side web browsers.

The possibility of leveraging independently operated search and index systems (e.g., Google) should not be underestimated, as it may actually reduce some of the pressure to build the search and indexer sub-components. By informing their indexing robots of the web archive's entry URLs (corresponding to source EPUs), a good deal of the web archive will become visible to external search engines. Moreover, if the archive gains a reputation for stability and acquires enough incoming links, archived objects would become increasingly prominent in external engine search results where document importance rises with the number of incoming links.

Candidate tools for the Archive Content Search

The Archive Content Search requires special-purpose tools. Although full text indexing and searching are well understood in some sense, user expectations change rapidly, and a wide range

of choices for scalable, open-source tools is lacking. While there is not a long record of experience with the tools appearing here, early results show promise. Candidate tools include:

- ❑ Apache Jakarta Lucene
<<http://jakarta.apache.org/lucene/docs/index.html>>
An open-source, high-performance, full-text search engine written in Java.
- ❑ Greenstone digital library building software from New Zealand
(<http://www.greenstone.org/>)
Constructs indexes that enable relevance-ranked searching of document texts and metadata using combinations of words and phrases.
- ❑ Guilda search engine <<http://dot.ucop.edu/guilda>>
A scalable, open-source search engine developed at the University of California.

5.6 Crawler

The Crawler service component is the software that downloads web-based materials. It would start by checking to see if the EPU has been crawled before and, if so, asks for confirmation before proceeding. Then it would download the EPU and, subject to collector- and archive-defined constraints, recursively follow hyperlinks extracted from the downloaded object. The downloaded materials would be re-assembled within the archive in a hierarchical manner that would mirror the source site(s). A log of every crawl (perhaps in summary form) would be kept for administrative purposes and accountability.

The crawler would be used by collectors to analyze the demographics of their intended collections before deciding how they should be captured and archived. Using the Crawler in this capacity, the collector would crawl the target materials, analyze the crawl once it is captured (using the same tools the Crawler provides to generate a report on the crawl), and then decide whether or not to keep the results of the crawl in the archive. Should a crawl prove unsatisfactory for any reason (e.g., it may have returned too much or too little material, or it may have downloaded files of the wrong type), the crawl results can simply be marked so the storage can be reused in a subsequent crawl.

As envisioned, the simple Crawler would support a single act of saving web-based materials to one centralized repository. Enhancements to this minimal service would include supporting subsequent and/or periodic crawls of an EPU that co-exist with prior crawls. Another enhancement would make the Crawler sophisticated enough to access the deep web, in other words, able to submit a comprehensive set of queries tailored to the search system behind a given HTML form, and to understand the search results. The enhancements have cost/benefit tradeoffs that would work differently for different organizations and audiences. Accordingly, it would be necessary to develop some criteria with which to prioritize development investments.

The Crawler would be operated via the Curator Interface that is described earlier. As already mentioned, the results of a crawl would be reported via email because of the potential duration of the crawl and the size of the response. For size considerations, the emailed response may itself be just a summary (e.g., number of objects downloaded, total time taken), with links to the mirrored content and a complete log, including errors, of the crawl's progress. Administration of the

Crawler service would call for computing simple statistics on a per-crawl basis and across all crawls.

Next to the downloaded materials, the most important by-products of a crawl are the metadata elements generated for each object that it captures. Among them are the original URL, crawl date, and HTTP headers that accompanied its capture. These elements seed the creation of object-level records that will be the basis for the object inventory within collections (different from the Registry, which is an inventory of collections). Before these records are used in searching and browsing operations, they will need, at a minimum, to be enriched by the assignment of persistent identifiers.

The crawl also generates other important by-products. It makes sense, for example, to initiate object processing as crawled objects are captured. The rationale is simple. Post-crawl object handling involves re-crawling the archive and is accordingly both slow and error-prone. What object processing is conducted will be contingent on the collector and their interests in building an archive of web-based materials. Two highly desirable object-processing options that would need to be implemented in the initial instantiation of the broker service include:

- ❑ full text indexing for those captured web objects that contain text—this enhances search and retrieval opportunities;
- ❑ computation and storage of object checksums or digests (a short string)—this supports de-duping for archives that periodically revisit web sites as well as content integrity validation.

Other object processing enhancements that could be made at crawl time are described briefly below.

Automatic metadata extraction creates new metadata elements extracted from downloaded documents. In the first instance, this would have to be a very crude heuristic procedure for recognizing such things as HTML title tags and meta-tags rather than an elaborate exercise in data mining.

Automatic text classifier generation creates new metadata elements containing text classifiers generated by text analysis algorithms. Initially, at least, we would envisage simple sets of machine-learning procedures for producing metadata that cost-effectively enriches the searching and browsing experience.

Establishment of relationships between crawls. This enhancement concerns the periodic re-harvesting of a set of web sites or, more generally, of a topic domain covered by a stable but not necessarily static set of web sites. The main issue with re-crawling is determining what relationship one crawl bears to the next crawl. The easiest option is to simply replace (destructively overwrite) the previous crawl. The next easiest response is to declare that each crawl stands as a separate collection on its own. But the problem is that patrons interested in the evolution of an object, site, or topic want to be able to focus on individual objects; this includes the appearance and removal of web site objects as well as the tracking of object changes over time. Moreover, a strategy has to be created to deal with searching and browsing mixed

collections of objects that exist at different times and in multiple related versions. It is no small challenge for collectors to define and implement relationships between old and new versions of a crawl.

Automated periodic recrawling. This feature would enable a collector to set up a crawl and have it automatically launched on a given EPU according to a regular schedule. The problems would be essentially the same as for manually initiated re-crawls. In particular, the main challenge will be for the collector to create the desired relationships (from the point of view of search and browse) between different crawls of the same material. In addition, because the potential number of different crawls could be large in an automated recrawl scenario, creating these sorts of relationships (if any) would have to rely more on automated means. With enough different version snapshots taken of an object over time, a group of snapshots could be viewed as a kind of "flipbook", or as a special kind of "object film" if the snapshots are spaced closely enough.

Candidate tools for Crawler

The Crawler component is another critical application that would require special-purpose tools. There are a wide range of commercial choices, but a much narrower range for scalable, open-source tools that have been recently reviewed. Candidate tools include:

- ❑ WebBase from Stanford University
<<http://www-diglib.stanford.edu/~testbed/doc2/WebBase/>>
Building on previous Google activity at Stanford University, this project aims to build the necessary infrastructure to facilitate the development and testing of new algorithms for clustering, searching, mining, and classifying web content.
- ❑ HTTrack open source offline browser)
<<http://www.httrack.com/>>
Allows users to download a web site to a local directory, duplicating the original site's relative link-structure. Fully configurable, it can also update an existing mirrored site and resume interrupted downloads.
- ❑ Teleport Pro offline browser
<<http://www.tenmax.com/teleport/pro/home.htm>>
A multi-purpose web spider capable of running up to 10 simultaneous retrieval threads, accessing password-protected sites, filtering files by size and type, and searching for keywords.
- ❑ An open-source, scalable web crawler sponsored by the Internet Archive
<<http://www.research.compaq.com/SRC/mercator/papers/www/paper.html>> Can be configured for different crawling tasks (such as collecting new statistics) by allowing users to provide their own modules for processing downloaded documents. Document finger-printing and URL normalization help avoid duplicate downloading and processing.
- ❑ Xyleme crawler developed at INRIA
<<ftp://ftp.inria.fr/INRIA/Projects/verso/gemo/GemoReport-229.pdf>>
Uses a ranking computation similar to Google's, but without storing the links matrix, which saves both storage capacity and time.
- ❑ Combine harvester from the DESIRE project
<<http://www.lub.lu.se/combine/>>
An open system for harvesting and indexing Internet resources. The parser reads the

harvested documents and extracts information such as links, headers, and metadata, and creates a catalog record that is written to the harvester database.

- ❑ ASByE hidden web searcher
<<http://www.lbd.dcc.ufmg.br/~debye/debye-family.htm>>
An agent-specification-by-example tool that generates a page collection plan by interacting with a GUI, providing examples of how to reach pages of interest, how to fill out query forms, and how to group related pages; this plan then feeds a programmable search agent.
- ❑ HiWE hidden web exposor
<<http://dbpubs.stanford.edu:8090/pub/2000-36>>
Parses an HTML form, computes the best (unused) values to submit with the form, confirms the submission did not yield an error response, and, finally, follows any hypertext links found in the response page.
- ❑ GNU Wget
<<http://www.gnu.org/software/wget/wget.html>>
A simple, open-source, non-interactive command line web mirroring tool that can run unattended (e.g., from scripts) and can resume aborted downloads.

Candidate tools for Metadata Extraction

The Metadata Extraction component is a special application area for which there is a limited range of tools. Software in this area is hard to build in ways that generalize easily, and what is available is often tailored to the needs of the developing organizations. The following tools have been favorably reviewed, but we expect to have to modify them significantly:

- ❑ DC-dot Dublin Core metadata editor from UKOLN
<<http://www.ukoln.ac.uk/metadata/dcdot/>>
Extracts and validates metadata from HTML resources and MS Office files.
- ❑ Mantis research toolkit from the OCLC CORC system
<<http://purl.oclc.org/mantis>> Used for building web-based cataloging systems with arbitrary metadata formats, definitions, and interfaces.
- ❑ DC Metadata Viewer-Constructor from the Chizhevsky Metadata Project
<<http://www.library.kr.ua/dc/lookatdce.html>>
An online form that tests a URL for pre-existing metadata and generates Dublin Core metadata for a page containing only traditional HTML metadata.
- ❑ ROADS project open source software from the UK
<<http://www.ilrt.bristol.ac.uk/roads/news/issue8/software/>> Software tools to help set up and maintain searchable, browseable subject catalogs of web-based materials.

Candidate tools for Automatic Classifier Generation

The Automatic Classifier Generation component is also a special application area for which there is a limited range of tools. Software in this research-oriented area generalizes fairly easily, but the results of its use may be hard to evaluate. This is an ambitious component to build out and tools are difficult to select. We expect to narrow the selection after gaining prototyping experience. Candidate tools include:

- SVMlight open-source software for Support Vector Machines
 <<http://svmlight.joachims.org/>>
 Fast, open-source software that applies to a range of machine learning problems in text classification and pattern recognition, and make efficient use of storage.
- Scorpion project from OCLC <<http://www.oclc.org/research/software/scorpion/>>
 An open-source automatic classifier for web-accessible text documents, Scorpion is for use with formal subject classification schemes or thesauri.
- The Matcher automatic classification tool from the DESIRE project
 <<http://www.desire.org/toolkit/matcher.html>>
 Performs subject classification by attempting to match extracted text words against a subject-specific thesaurus, and applying some heuristic processing to the results.
- iVia open-source subject portal system from INFOMINE
 <<http://infomine.ucr.edu/iVia/>>
 Includes a PhraseRate program that allows fully- or semi-automated extraction of significant words and phrases from a document; this optionally feeds the automatic assignment of Library of Congress subject headings (using a k-nearest neighbor classifier that relies on expert-vetted training sets), which in turn feeds the assignment of Library of Congress classifications.
- Klarity automated text analysis tools
 <<http://www.intology.com.au/20products/50Klarity>>
 Performs concept-based categorization, keyword extraction, and automatic summarization of web pages. "The output can be customized to various formats including meta-tags, RDF, and ASCII output (for upload to databases)."

5.7 Indexer

The Indexer service component is divided into two parts, corresponding to the registry and the archive content search components, respectively. In support of object-level search, the Archive Content Indexer is invoked at crawl time. In support of collection-level (EPU-Registry-level) search, the Registry Indexer is invoked when the registry is updated. Both parts must be capable of creating indexes that support fielded search and search for text inside a field.

The Archive Content Indexer must, at a minimum, be capable of parsing the HTML file type. Given the prevalence of PDF and MSWord file types, it should also be able to parse these types.

An important requirement of the Archive Content Indexer will be its ability to add index entries incrementally and to a very large scale. The archive is expected to grow rapidly and continually, and incremental indexing avoids the resource-intensive rebuilding of indexes from scratch.

Candidate tools for Indexer

The tools in this section are nearly identical to those for the Registry and the Archive Content Search components, since indexing tools are normally packaged with search systems. One class of tools specific to this area is format parsers and converters, which are critical to the ability to “see inside” a document. Here, we list only one tool because it handles the most important non-transparent format that we expect to receive:

- ❑ Open-source XPDF Portable Document Format viewer
<<http://www.foolabs.com/xpdf/>>
the PDF to text utility in the XPDF distribution parses a PDF into indexable text words.

6 Sustaining the broker service

Safeguarding U.S. governments' web-based information will require widespread effort by the libraries and other memory organizations who are, for a variety of historic and other reasons, ideally suited and eager to undertake the task. Highly centralized efforts, though valuable in their own right, are unlikely to capture all of the web-based materials that users might be interested in, or describe and organize the information in a meaningful way for the many communities that will be interested in using it. There is a need to encourage memory organizations to take up their historic roles in a cost effective manner. Critical here is the supply of configurable tools and deep technical infrastructure that enable archives to select and capture materials they are interested in, and to describe and organize those materials in ways that meet local needs. For the supply of common tools and infrastructure, we have looked to a broker service and have defined what that service might look like at the CDL. The following section reflects on the means of financially sustaining a broker service. This is not a formal business plan; rather it is an investigation into costs and the revenue streams that may offset them.

6.1 Cost elements

Development of the basic infrastructure

The CDL estimates that an initial instantiation of the tool suites described above, including a early working version of the repository and a compilation of materials that will support its use, will require a development effort costing \$1,500,000 over two years. The estimates understate the actual cost, since development will leverage investments already being made by the CDL and its partners:

- ❑ The University of California Berkeley Library and the San Diego Super Computer Center will leverage local computer infrastructure to supply redundant data storage and data backup.
- ❑ The San Diego Super Computer Center has invested considerably in the development of its Storage Resource Broker, which will emerge as an integral component of the CDL's technology infrastructure.
- ❑ The Stanford University Computing Science Department has invested in the development of its WebBase crawler.
- ❑ The CDL has developed a number of essential technical components for the ingest, encoding and representation of archival objects, persistent naming, and the search and retrieval of archived objects.

Ongoing development of the basic tool suite

The development described in section 5 indicates an initial configuration for the tool suites as well as possible enhancements. The initial broker service will not include enhancements, which will need to be constructed as part of ongoing development. In addition, we anticipate needing to refine and improve the tool suites in response to their initial use (both by collectors and the broker). Ongoing development will cost \$200,000 – \$230,000 *per annum*, which is the cost associated with two programmer analysts, one who would be at a senior level and devoted full-time to this endeavor.

Basic low-level support for data capture

Crawlers are difficult to manage. Crawls may fail for any number of reasons and need to be restarted. They can also go astray, capturing more and different content than the collector intends. Even where collectors assume responsibility for managing their crawls, some low-level support will be required at the broker service to monitor and help efficiently queue crawls, and to ensure availability of adequate storage and processing capacity. Effort required by the broker will extend to a person or persons responsible for knowing what crawls are in process, what crawls are in the queue, and who will monitor crawls that are in progress. How many FTEs will be required in this role will depend upon the number of collectors who are using the broker service, and the extent and frequency of their crawls.

Support for collecting institutions

We anticipate needing to provide a high-level of support to collecting institutions in scoping collections and determining how best to capture, organize, and describe the web-based materials they will comprise. Some support will be available from web-based guidelines. Still, since we want to encourage collectors to work with different (though possibly overlapping) sets of web-based materials and capture and manage these materials in different ways, we anticipate needing to supply a high level of personal service in a consulting role. In this role, the broker will inform collectors about what others are doing and give advice, based on their experience with the strengths and weaknesses of the different approaches. Again, the level of effort that is required will depend on the level of use that the broker service generates. Particular emphasis is likely to be given to institutions that are initiating collection development activities with the broker services.

Persistent management of archived collections

In the service model described here, data are stored in the repository managed by the broker, or in third-party repositories selected by the collecting institution. In the latter case, financial responsibility for preservation rests outside the broker's purview. In the former case, the costs can be distributed since the broker will preserve the numerous collections that are managed in its repository as if they made up a virtual uniform collection. Accordingly, the costs of preserving any one collection will be a portion of the total cost of preserving all of the collections that are managed by the broker.

Access to collections

The broker will enable a basic level of access to archived collections as described above in section 5 and will require little ongoing investment in filestore, processing and network capacity, as well as modest service upgrades. More fully featured access and end-user services, whether

they are built for one collection or for a number of discrete collections, will require additional investment at a level commensurate with the nature and sophistication of the intended service.

6.2 Revenue streams

Permanent institutional support

If located at the CDL, a broker service would benefit from a substantial level of permanent institutional support. The UC libraries need to collect in areas (such as government information and many area and other studies) where key information resources are largely available in web-based formats. To do this, they are investing in the necessary infrastructure—an infrastructure suitable to web archiving—and accordingly, can be leveraged to serve a broader web-archiving community. Admittedly, the initial set-up costs of \$1,500,000 may be problematic; the libraries are likely to seek assistance through external grant funding. Such funding promises to speed the development process. It will also help ensure that parochial efforts reflect, inform, and integrate with complementary activities underway elsewhere in the U.S. and abroad. Ongoing support costs are more easily met than start-up ones by permanent institutional support, as they will leverage existing organizational capacity as indicated briefly below.

- The UC libraries' preservation program. The program is supported by and maintained on behalf of the UC libraries. Comprising only a small number of FTEs, the program focuses on the strategic direction, management, and oversight of preservation initiatives. It integrates work of the CDL with its strategic partners and will act in a business development capacity by building relationships with libraries and other organizations that seek to use the CDL's broker services as a means of building persistent collections of web-based materials. For operational and technical capacity, it relies upon other operational service units within the CDL as listed below. In support of the web archiving activities described above, members of the preservation program will be responsible for business development, liaison with collectors and other partners, and supplying user support in the consulting role defined above.
- CDL technologies. The technologies group is the largest unit in the CDL, and is responsible for the development and maintenance of technical infrastructure as required to support the full range of CDL services. The unit includes five groups, all of which will be leveraged in support of the broker service indicated here.
 - The Advanced Technology group will lead the research and development effort necessary to design, build, and enrich the tool suites identified previously. It is also responsible for the development and maintenance of the CDL's digital archival repository and for the persistent management of its contents.
 - The Ingest and "Quality Assurance groups were created to support the CDL's union bibliographic catalog, the Melvyl Catalog, and to acquire and process the 350,000 MARC records CDL receives every week for the catalog. The groups were recently extended to support acquisition and data processing for all of the CDL services that acquire and manage digital information content. These units will provide low-level support for the crawler as described above.

- The Access Services group comprises a pool of programmers whose members will be available for work on development, maintenance, and enhancement of the basic infrastructure and tool suites.
 - The Architecture and Infrastructure group assures a coherent and integrated systems environment for CDL services. Its purview extends across hardware and software maintained by CDL and third-parties.
- Digital Library Services. This group is responsible for the online services that organize, integrate, and present the digital collections managed by CDL. Those collections extend to some 250 databases, 9,000 online journals, the 25,000,000-record Melvyl Union Catalog, 5,000 statistical databases, a collection of nearly 200,000 digital images, a union catalog of some 7,000 finding aids for UC archives and special collections, and the content of the eScholarship Repository. With existing services, the unit maintains and continuously evaluates these collections, extending and enhancing services in light of their evaluation. This division also identifies new service needs, builds functional specifications, and conducts product and technology reviews as appropriate. Regarding the preservation effort, the unit will provide leadership in the specification, evaluation, and development of the tools and services that:
- present preservation capacity to client libraries through an administrative interface;
 - enable basic resource location, discovery, and retrieval for archived information; and
 - enable the federation and aggregation of materials managed in distributed digital archives.
- Education, outreach, and assessment responsibilities are self-explanatory and will play an important role. Through assessment, the CDL identifies appropriate evaluation methodologies for new and existing services, creates evaluation tools, analyzes quantitative and qualitative data, and makes recommendations to service managers. The education and outreach provides frontline support to libraries and other information organizations that use CDL services and tools. With regard to the web-archiving efforts, the unit will be responsible for evaluating collectors' needs and how the tool suites will support them. Its work will be critical in defining and prioritizing desirable service enhancements. The unit will also play a part in the business development and consulting functions that are described above for the broker.

Although permanent institutional support can support the broker service for use by UC libraries and other memory organizations, additional funding will be required to meet the costs involved in making the service available to collectors outside UC libraries. Revenue sources are listed below and are likely to be required in some combination. Revenue sources are grouped into those that may originate with collectors (who may be expected to pay to use certain broker services) and with users who access archived collections.

Revenues from collectors

Collectors outside the UC system may be charged for their use of specific broker services as described below. The total cost to collectors, however, cannot be so high as to discourage their

involvement in web-archiving activities. Our strategy for keeping costs low is to ensure that the sums paid by collectors leverage their own curatorial expertise as well as the permanent and other sources of investment available to the broker.

Collection design (decisions about collection identification, capture, metadata enrichment, and access strategies, for example) will be undertaken by specialists working within collecting institutions. The effort will require specialists at the collecting institution to work in close consultation with those who are able to speak for targeted user communities and their needs. It may also require support from staff at the broker service that are able to provide insight into what other collecting institutions are doing and guidance about the strengths and weaknesses involved in different approaches. The CDL may charge collecting institutions for this consulting service, probably at a daily rate. Alternatively, the CDL may work under contract with the collecting institution to design collection strategies that meet the collecting institution's needs. Acting in this role, the CDL would charge the collecting institution at a level commensurate with the size and complexity of the collection and the number and nature of interactions required with the collecting institution's staff and its users.

Configuring and implementing capture routines, and validating, naming, and enriching crawled content with resource discovery and other metadata. Here, too, the burden will fall largely on the collecting institution as their staffs use the broker's tool suites. Although support for the use of these tools will be available through freely accessible online materials, a charge may be made for more personal assistance. Charges would be made at a daily or hourly rate. Alternatively, the CDL may act under contract to capture, organize, and describe the collection the institutions require. Here, charges would be set at a level commensurate with the size, complexity, and frequency of the collecting task. Regardless if the collecting institution governs its own crawl(s) and enriches the captured content, or contracts these processes out to the CDL, the CDL would also charge the collecting institution to recover the costs of the processing and filestore capacity required.

Preservation. As indicated above, content captured by collecting institutions may be stored in a repository managed by the CDL or by a third party. If stored at CDL, the collecting institution would pay a sum comprising the ongoing cost of storage and an annual contribution toward the cost of the data's periodic migration.

Access. The CDL would make collections available via simple browsers as described above and may seek to recover the cost of the filestore, processing, and network capacity from collecting institutions. More fully featured access services are possible, but they would have to be developed on an *ad hoc* basis for a fee. Since the development of tailored access services may distract the CDL's attention from essential core functions (such as developing and maintaining the repository and tool suites, and supporting collecting institutions in their use), it is likely to undertake the development of highly tailored access services only under exceptional circumstances.

Revenues derived from user of collections

The terms and conditions of use for any of the archived collections that are built with the broker's assistance will be established by the collecting institution in consultation with the CDL. In all cases, the collecting institution will be responsible for safeguarding any intellectual property and copyrights that are bound up in their archived holdings. Collecting institutions will also be responsible for determining whether or not access fees apply to their holdings. Where such fees do apply, the development of appropriate mechanisms to register, authorize, authenticate, and collect fees from users will need to be undertaken in consultation with the CDL. In support, the CDL would seek to supply a range of enabling tools. In most cases, the CDL would expect them to be configured, applied, and administered by the collecting institution. Any revenues derived from the use of collections would return to the collecting institution.

As archived collections grow in number and in size, opportunities may arise for the CDL to develop a range of end-user services that integrate access to numerous underlying collections. For example, it may be possible to selectively integrate materials residing in several archives to present thematic collections, or to build portal services that integrate archived holdings with other online information resources (as may exist, for example, in commercial journals, digital library collections, or even other archives). These services could add very substantially to the value of archived collections and provide opportunities for the CDL to financially sustain itself in its broker role.

Revenues derived from use of the tool suites

The CDL will seek wherever possible to build open-source tool suites and make them and the underlying code available at no cost.

Conclusion

The broker service is likely to be financially sustained through a variety of means, including core institutional investment, revenues derived from users, and revenues derived from the use of any value-added services the broker is able to build on top of the archived collections it manages. Given the level of investment that will be required by the broker, and the importance of minimizing costs to memory organizations, broker services may be best suited to institutions that are able to extensively leverage existing investments—that is, in institutions that are developing web archiving capacity to fulfill their own local missions.

7 Web Archiving In Context

In a survey of digital preservation, historian Roy Rosenzweig suggests that his professional colleagues should become more directly involved in the activities of archivists and librarians:

Historians need to be thinking simultaneously about how to research, write and teach in a world of unheard-of historical abundance and how to avoid a future of record scarcity. Although these prospects have occasioned enormous commentary among librarians,

archivist, and computer scientists, historians have almost entirely ignored them. In part, our detachment stems from the assumption that these are ‘technical’ problems, which are outside the purview of scholars in the humanities and social sciences. Yet the more important and difficult issues about digital preservation are social, cultural, economic, political, and legal—issues that humanists should excel at. The ‘system’ for preserving the past that has evolved over centuries is in crisis, and historians need to take hand in building a new system for the coming century.⁴⁰

Rosenzweig’s article affirms the diversity of information types required by students of society, politics, and history, and that the preservation patterns of the analog world must be reinvented in the digital environment.

Capturing and preserving web-based government publications is just one element in a broader strategy whose goal is to preserve a significant record of government activities for use citizens and scholars now and in the future. A holistic approach to the preservation of all permutations of government information requires a multifaceted strategy. Specifically, tactics must be developed to address the preservation of a broad range of government information categories, including:

- ❑ print publications and archival sources;
- ❑ digital content on portable media (tape, diskette, CD-ROM, DVD, PDA, etc.);
- ❑ databases (geospatial, numeric, etc.);
- ❑ electronic administrative records (email, e-government transaction records, etc.); and
- ❑ software applications.

The development of a holistic approach to preserving the record of government is beyond the scope of this report. However, it is important to recognize the substantial challenges that remain in other areas. Solutions to the web-archiving problem are important, but they are only partial solutions in light of this far broader challenge.

New thinking about traditional print materials

Many large government publication collections are held in state libraries and archives. Most are chronically underfunded. Recent budget crises within state governments have exacerbated this trend, and in some cases state libraries and archives are faced with comprehensive funding cuts that effectively eliminate their preservation programs. Recent examples include the 60 percent cut to the budget of the Washington State Library and Governor Jeb Bush’s proposal to eliminate the Library of Florida.⁴¹ Due to these funding cuts, libraries are increasingly ending their participation in depository programs that provide a framework for systematic collecting and preservation. Many government collections, including valuable older collections, are located in regular library stacks. They are not maintained according to the standards for security, preservation, and climate control typically applied to other special collections.

⁴⁰ Roy Rosenzweig. Scarcity or Abundance? Preserving the Past in a Digital Era. *American Historical Review*, June 2003. pp. 736.

⁴¹ Darlington, David., State History Programs in Crisis. *Perspectives Online: Newsmagazine of the American Historical Association*, April 2003. <http://www.theaha.org/perspectives/issues/2003/0304/0304new1.cfm>

Interestingly, while our investigation into web archiving suggests a strong and compelling reason to support varied and highly redundant curatorial efforts, recent thinking about the management of duplicate print materials is moving in an altogether different direction. In light of the budgetary constraints indicated above, the GPO, for example, is beginning to question its approach to preserving government print publications as represented in its Federal Depository Library Program. In effect, it is asking whether the persistence of government print publications really requires numerous and highly redundant print collections, each of which contain an overwhelming number of infrequently used items.⁴² The University of California, which houses eight FDLP libraries (not including UC's four federal depository law libraries) is questioning its print preservation policies along the same lines. A task force was recently convened to evaluate prospects for building shared collections of some government materials, and recommended that a shared print repository be established so individual FDLP libraries within the UC system could focus on unique materials that were directly related to their local academic programs and community interests without sacrificing access to common materials that were beneficial to all.⁴³ The report also notes that the transition from print to a web-based distribution of government information has created a challenge for reliable long-term access to this important resource category:

To a much greater extent than in other disciplines, where digital resources are often digital editions of well-distributed and well-indexed paper publications, government information in digital formats is characterized by poor bibliographic access, and is highly volatile. Government information web sites change frequently, especially with the change of an administration or the reorganization of an agency. There are a number of federal initiatives, both governmental and partnership to preserve digital titles, however, these initiatives are small and localized and do not begin to address the need for a systematic plan for the preservation of digital government information. Of particular concern to UC librarians are regional Federal publications (forest plans, environmental impact statements, for example) in California and the Pacific North and South West. (pp. 5–6)

A recent report published by the Council on Library and Information Resources tilts the same way, though with respect to a far broader range of print materials than that produced by federal, state, and other governments. In "Developing Print Repositories" Bernard Reilly Jr. formulates an argument that begins to suggest that print preservation of non-unique materials is a service that might most effectively be provided as a shared utility whose support is shared by interested libraries. The model's advantage is that it pools scarce resources in the provision of a common good while freeing up libraries to invest more in persistently managing holdings that are rare and unique.⁴⁴

⁴² See remarks by Judy Russell, U.S. Superintendent of Documents. *Future Directions of the Depository Library Program*, at the 142nd Association of Research Libraries Membership Meeting, Federal Relations Luncheon, May 15, 2003. <http://www.arl.org/arl/proceedings/142/russell.html>

⁴³ Final Report, Task Force on Government Information, University of California Shared Collection of Government Information. Submitted to the UC Systemwide Operations and Planning Advisory Group. May 2003. http://www.slp.ucop.edu/sopag/govinfo_finrept.pdf

⁴⁴ See Reilly, Bernard J Jr. *Developing Print Repositories: Models for Shared Preservation and Access*. (CLIR, 2003) at: <http://www.clir.org/pubs/abstract/pub117abst.html>

Portable digital media and outmoded hardware and file formats

Government information collections possess a large number of digital materials distributed on highly volatile handheld media including: CD-ROM; DVD; and 3.5-inch and 5.25-inch floppy disks. One indicator of the size of these collections is the number of titles distributed by the Government Printing Office in this format between 1995 and 2002. A summary is provided in Table 6 below.

The challenges with these materials are well known and have been written about frequently. The media themselves are subject to rapid physical degradation. Key causes of the limited lifespan for these media include the risk of exposure to magnetic fields and internal deterioration processes, which destroy the stored data. Although optical storage media should last much longer than magnetic media, they are also subject to problems. CD-ROMs must be stored and handled carefully and even so, might not last longer than, say, 10 years. The hardware and software that are required to read these media become obsolete even more quickly. And with the possible exception of ASCII, the file formats that are used to record information on these media – especially the proprietary formats – are also at risk.⁴⁵

Despite all this, there has not to date, been any systematic approach to the persistent management of these materials. They are consequently substantially at risk.

Table 6. Number of titles distributed by the GPO in handheld digital formats, 1995-2002

Portable Digital* Titles Distributed to Federal Depository Libraries, 1995–2002		
Fiscal Year	Media**	Number of Titles Distributed
FY 2002	CD-ROM, DVD	483
FY 2001	CD-ROM, DVD	480
FY 2000	CD-ROM, DVD	617
FY 1999	CD-ROM, DVD	682
FY 1998	CD-ROM	836
FY 1997	Electronic	741
FY 1996	Tangible Electronic	639
FY 1995	Tangible Electronic	412

*Portable digital includes CD-ROM, DVD, and 3.5-inch and 5.25-inch floppy disks. **Media names reflect the GPO's language. Media distributed in FY 1995–1997 may have been CD-ROM and/or floppy disk.

Databases

Government agencies typically gather vast stores of information from and about citizens, businesses, economic life, governments, and non-governmental organizations. The data is used to support agency activities, produce reports for government officials and the public, and fulfill statutory and regulatory obligations. Enormous databases are created from these data-gathering and compilation activities. While it may be possible to develop sophisticated crawlers that can interact with and extract content from some of these databases in their web-iteration, the source databases are typically managed behind impermeable network walls. Preservation of their content must be accomplished in collaboration with the host agency. The European Resource Preservation and Access Network (ERPANET) held a three-day workshop on the “Long-Term Preservation of Databases” in Berne Switzerland April 9–11, 2003. The presentations made in

⁴⁵ cf Ross, Seamus and Ann Gow, *Digital Archaeology. Rescuing Neglected and Damaged Data* (JISC, 1999), 1-16. <http://www.ukoln.ac.uk/services/elib/papers/supporting/pdf/p2con.pdf>

that forum showed there is significant activity within a number of government agencies addressing the long-term preservation of these databases.⁴⁶

Administrative records

Perhaps the most challenging arena for preservation will be the host of issues raised by the preservation of administrative records that document the internal activities of government agencies, such as correspondence within agencies and correspondence between agencies and citizens. In the analog environment, this meant developing a standard retention policy and transferring standards to the central government archive or the agency archive. With the advent of “e-government,” these sets of practices are difficult to re-establish. The range of e-gov activities is large and growing.⁴⁷ Examples include:

- ❑ electronic mail;
- ❑ online submission of tax forms and other required reports;
- ❑ online applications for passports, social security, vehicle registration, and other goods and services; and
- ❑ participation in voting (teledemocracy) and rule-making (e-regs).

Through its electronic records program, the National Archives and Records Administration is making great strides in addressing the challenge of electronic records management. The NARA program, however, concentrates largely if not exclusively on federal government records. Whether and to what extent the program’s good work can be taken up or even influence complementary efforts at state and local levels remains to be seen.

Software applications

Government agencies routinely develop digital resources on the basis of commercial and open-source software applications. Occasionally, government agencies develop software applications internally. Although theoretically abstract from the associated content, these applications are often required for the exploration, extraction, and presentation of the content. As such, any strategy for long-term access to government information—indeed any category of information—must address the preservation of software applications.

⁴⁶ For example, see the conference briefing paper available at:
http://www.erpanet.org/www/products/bern/erpaWSBern_Documentation.pdf

⁴⁷ For an overview of the range of issues involved, see Digital Government: New digital tools, projects, and resources are making government more accessible to its citizenry. *Communications of the ACM*, v.46: 1. January 2003.

8 Conclusion

Institutions such as libraries, museums, and other memory organizations working with digital content are facing some common problems. They all have encountered a range of large-scale, long-term curation and preservation challenges that strain their local resources. As a result, project managers are typically looking beyond their own organizations for ways to organize, present, and preserve digital information.

However, due to the diverse needs of each institution, no single model fits each one's needs. Organizations need a variety of capture and curation tools that allow them to respond to the specific needs of their local constituencies.

Adding to the problem is the complexity of the dot-gov domain. Our analysis indicates that this domain is large and constantly expanding, volatile (with constant content addition and removal), and opaque. In addition, any solutions for the curation and preservation of web-based content need to incorporate solutions for preserving the full range of government information, including other digital formats and analog content.

A service model that distributes responsibility for different parts of the preservation process to organizations that are best suited to undertaking that responsibility promises to address the problems of preserving web-based content, with tools for data capture and analysis, preservation and storage, curation, and administration. A broker service is an essential part of this service model and will supply a range of tools and a deep technical infrastructure that will enable memory and other organizations to focus in areas where their expertise and their missions naturally permit.

Through our work we have identified some of the key roles that the broker service must play to encourage memory organizations to take a more widespread and active role in preserving web-based government materials. It has also helped us identify the means by which the broker service may sustain itself financially while keeping costs down for memory organizations that wish to take up an web archiving role.

Appendix 1: Question Sets and Individuals Consulted

Interview Question Set A: Overview of Initiatives

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 dot-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?

Interview Question Set B: Project Level Interview

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; and
 - 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?

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
UCLA
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
UCLA
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 2: Projects and Programs that Informed our Research

United States Federal:

1. CyberCemetery
 - a. <http://govinfo.library.unt.edu/default.html>
 - b. Principals: University of North Texas/GPO Partnership Agreement
 - c. Initiated: 1997
 - d. Scope: web sites of defunct U.S. federal agencies
2. DOSFAN
 - a. <http://www.uic.edu/depts/lib/documents/resources/dosfan.shtml>
 - b. Principals: University of Illinois, Chicago; U.S. Department of State, GPO Partnership Agreement
 - c. Initiated: 1993
 - d. Scope: digital U.S. State Department documents
3. Electronic Records Archive
 - a. http://www.archives.gov/electronic_records_archives/index.html
 - b. Principals: NARA, San Diego Supercomputer Center
 - c. Scope: electronic records, which may include web sites; the long term preservation of various formats, including PDF, image formats, HTML, etc.; issues of collections versus items within a collection; looking at new partnerships with agencies to share the responsibility and develop more advanced tools
 - d. Presentations: several presentations are available via the web
4. Minerva, the Web Preservation Project
 - a. <http://www.loc.gov/Minerva>
 - b. Principal: Library of Congress
 - c. Initiated: 2000
 - d. Scope: 2000 and 2002 elections web sites and September 11, 2001 collections
5. 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 are made web-accessible through document management systems) within the NASA GSFC domain; they are starting a research phase to determine the most efficient way to capture and store the GSFC web pages
6. National Digital Information Infrastructure and Preservation Program (NDIIPP)
 - a. <http://www.digitalpreservation.gov/index.php>
 - b. Principal: Library of Congress
 - c. Initiated: 2000
 - d. Scope: developing strategies for digital preservation

7. National Library of Medicine
 - a. Permanence Rating System
 - b. Principal: National Library of Medicine
 - c. Scope: metadata and permanence rating system for NLM web sites and electronic publications
 - d. Reports: Phase II Report from the Permanence Ratings Committee ><http://www.nlm.nih.gov/pubs/reports/permanence.pdf>); fact sheet on preservation/electronic resources ><http://www.nlm.nih.gov/pubs/factsheets/preservation.html>)
8. National Technical Information Service
 - a. NTIS Science Portals Program
 - b. Principals: National Technical Information Service/Department of Commerce
 - c. Scope: harvesting documents from several science agencies' web site, including the Department of Energy, and archiving for distribution to the public
 - d. Presentation: <http://www.science.gov/workshop/wfinch.pdf>
9. 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
10. USDA Digital Publications Preservation Program
 - a. <http://www.nal.usda.gov/preserve>
 - b. Principals: National Agricultural Library and the USDA Economic Research Service
 - c. Scope: metadata, metadata template, framework document for the preservation of USDA digital publications (this may or may not classify as government web sites)
 - d. Reports: framework document and other reports available via the web site
11. 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. Scope: digital content
12. Web Documents Digital Archive Pilot Project
 - a. http://www.niso.org/presentations/barnum-ppt_01_22_02/
 - b. Principals: U.S. GPO in cooperation with OCLC
 - c. Initiated: 1999
 - d. Scope: electronic U.S. federal documents

State and Local:

13. 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
14. 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
15. 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 web sites; retrospective digitization of campaign literature
16. 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 web pages
 - e. Report: http://www.isrl.uiuc.edu/pep/papers/UIUCLIS_2001_9_EARCH.html
17. Texas Electronic Depository
- a. <http://www.tsl.state.tx.us/lot/electronicdepositorylib.html>
 - b. Scope: electronic documents
18. 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
19. 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 the OCLC Web Preservation Project
 - c. Initiated: 2001
 - d. Scope: web-only state publications and records

Foreign:

20. ARCHIPOL

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

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

- a. <http://www.antwerpen.be/david/eng/index.htm>
- b. Principals: Max Wildiers Foundation—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

22. Kulturaw3

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

23. NEDLIB - Networked European Deposit Library

- a. <http://kb.l/coop/nedlib>
- b. Principal: The Koninklijke Bibliotheek, National Library of the Netherlands
- c. Initiated: 1998
- d. Scope: technical infrastructure

24. netarchive.dk

- a. <http://www.netarkivet.dk/index-en.htm>
- b. Principal: Denmark's Electronic Research Library
- c. Scope: tools and collecting strategies

25. 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

26. Our Digital Island

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

27. 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

28. WARP

- a. <http://warp.ndl.go.jp/>
- b. Scope: tools

WebArchiv

- c. <http://webarchiv.nkp.cz>
- d. Principal: National Library of the Czech Republic
- e. Scope: preservation of and access to Czech resources

Memory Organizations:

29. Infomine

- a. <http://infomine.ucr.edu>
- b. Principal: UC Riverside
- c. Initiated: 1993
- d. Scope: metadata generation and indexing

30. Lots of Copies Keeps Stuff Safe

- a. <http://lockss.stanford.edu/>
- b. Principal: Stanford University
- c. Initiated: 1999
- d. Scope: persistent digital caches of HTTP-delivered content

31. Political Communications Web Archiving Project

- a. <http://www.crl.edu/content/PolitWeb.htm>
- b. Principal: Center for Research Libraries
- c. Initiated: 2002
- d. Scope: develop effective methodologies for the systematic, sustainable preservation of web-based political communications

Not-for-Profit Organizations:

32. Internet Archive

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

33. 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

For-Profit Organizations:

34. Google

- a. <http://www.google.com>
- b. Scope: search engine, metadata, and content

35. Lexis-Nexis/Reed-Elsevier

- a. <http://www.reedelsevier.com>
- b. Scope: metadata and content

36. Newsbank

- a. <http://www.newsbank.com>
- b. Scope: metadata and content

37. Rand California

- a. <http://ca.rand.org/stats/statistics.html>
- b. Scope: statistics

Appendix 3: Sample Models for Capture, Curation, Preservation

Model	Description	Advantages	Disadvantages
Model 1 Click, print, bind	Locate digital content; capture it, print it, and then handle it using existing processes for print publications.	<ul style="list-style-type: none"> ▪ Libraries know how to manage print publications—allows libraries to leverage existing selection, cataloging, etc., workflows. ▪ Libraries understand the costs and actions associated with the long-term preservation of paper materials. ▪ Might be cheaper to preserve. ▪ No added costs to deliver items to users. ▪ Can support local needs (e.g. reference tools). ▪ Resists change/increased authenticity. ▪ Easy, low-tech solution. 	<ul style="list-style-type: none"> ▪ Since this process is not guaranteeing access to the digital object, online remote access may be limited. ▪ Functionality loss in conversion from digital to analog (keyword searching; absence of an index, etc.). ▪ Captures only a particular view or version ▪ Heavily manual process, very selective, and will miss a lot. ▪ Severely limits the type of materials. Will not work for materials that don't "live well" in an 8 ½ x 11 format. ▪ Might incur additional cataloging costs. ▪ Shelf space/stacks is expensive.
Model 2 Click, save to disk	Locate digital content; capture it, save it to a local machine. Once saved, decisions about storage, access, and preservation must be made.	<ul style="list-style-type: none"> ▪ Original functionality can be maintained for certain types of digital objects ▪ Provides an interim solution—in the absence of a digital preservation strategy, it can buy time. ▪ Libraries might be able to leverage existing workflows (cataloging, etc.). ▪ Intermediaries have control over what they want to capture. 	<ul style="list-style-type: none"> ▪ Not scalable. ▪ Will only work with specific types of digital materials. ▪ Difficult to preserve the full functionality of an entire site. ▪ Demands storage space. ▪ Demands a digital preservation strategy (migration, emulation, etc.). ▪ Demands an access strategy (must determine how to integrate with other services, and what type of metadata).
Model 3: Targeted crawl (automated)	Capture content according to particular criteria—for example, all files modified after a certain date, with particular metadata, in a particular format.	<ul style="list-style-type: none"> ▪ Requires less human intervention than Models 1, 2, and 4. ▪ Can take advantage of intellectual input and knowledge of intermediaries -- an intermediary can set criteria to crawl all USGS w/ "California" in the title. ▪ Can be carefully tied to collection development policies. 	<ul style="list-style-type: none"> ▪ Subject to the constraints of the crawler. ▪ The more automated the approach, the less precision?

Model	Description	Advantages	Disadvantages
Model 4: Interactive crawl (facilitated)	Initiate a crawl. As the crawl proceeds, reports are issued requiring input. The crawler is prompted to accommodate this input; this process is repeated as the crawl proceeds. As the crawler encounters a problem, an event log is created for human review. For example, the crawler encounters a form, prompts a human for help, the human fills in the form, and then the crawler captures the site.	<ul style="list-style-type: none"> ▪ Might assist in highlighting areas of the deep web for intermediaries to make a selection judgment. In some cases, those pages may be transactional pages that are not documents. ▪ If crawled materials are not appropriate, (such as transactional pages), they can be disregarded ▪ Can take advantage of intellectual input and knowledge of intermediaries. 	<ul style="list-style-type: none"> ▪ Demands that an intermediary “baby-sit” the crawl. ▪ Question of scalability. ▪ Question of ability to develop technology.
Model 5: Negotiate with producer	Negotiate with the data producer to acquire content (potentially web and non-web content).	<ul style="list-style-type: none"> ▪ Producer may have a great deal of information that is not publicly accessible on the web. A library can negotiate to acquire these materials. ▪ May provide the ability to secure deep web materials. ▪ Get the underlying data and build your own interface/search engine (original interface was not good). ▪ Might allow libraries to dictate to the producer how we want the materials (metadata, standards, etc.). ▪ May have a positive impact on the authenticity question. ▪ Once the library has the data, it can be better integrated into other services. 	<ul style="list-style-type: none"> ▪ Question of scalability. ▪ Requires a large amount of resources to negotiate arrangements with producers. ▪ Negotiated arrangements might be dependent on a personal relationship or contact. ▪ This method leaves us beholden to the producer.

Model	Description	Advantages	Disadvantages
Model 6: Save blindly	Perform global capture without prior selection or analysis activities, other than targeting a domain or set of sites.	<ul style="list-style-type: none"> ▪ Useful as an emergency measure—get a lot of stuff before it goes away. ▪ Buys time. ▪ Bulk crawl can serve as a safety net (CRL project) —develop a strategy for capture/preservation, but keep the crawl just in case. ▪ Simple programming—several crawlers already perform in this manner. 	<ul style="list-style-type: none"> ▪ Requires large amount of disk space. ▪ Need for tools to help generate metadata. ▪ Winnowing wheat from chaff may be harder (more chaff). ▪ If there is no metadata, the context of where the web page came from becomes more important (not only where it's stored, but also the pages that link to it). ▪ Big crawls take more time and space.
Model 7: Save blindly, then select	Perform global capture. Select materials on the stored content, ex post facto.	<ul style="list-style-type: none"> ▪ Buys time. ▪ Bulk crawl can serve as a safety net (CRL project) —develop a strategy for capture/preservation, but keep the crawl just in case. ▪ Can feed back into many of the models. 	<ul style="list-style-type: none"> ▪ An enormous amount of unusable material will be captured. ▪ Requires the development of tools for access. ▪ Too many materials to catalog—materials won't appear in OPACs.

Appendix 4: List of Sources for Service Vision

Journal Articles

Arms, William Y et al. Collecting and Preserving the Web: The Minerva Prototype. *RLG DigiNews* 5:2(2001).

<http://www.rlg.ac.uk/preserv/diginews/diginews5-2.html#feature1>

Ipeirotis, P. G., L. Gravano, and L., M. Sahami. Probe, Count, and Classify: Categorizing Hidden-web Databases. *Sigmod Record* 30:2 (June 2001):67-78.

Masanès, Julien. Towards Continuous Web Archiving: First Results and an Agenda for the Future. *D-Lib Magazine* 8:12 (December 2002).

<http://www.dlib.org/dlib/december02/masanes/12masanes.html>

Mitchell, Steve et al. iVia Open Source Virtual Library System. *D-Lib Magazine* 9:1 (2003).

<http://www.dlib.org/dlib/january03/mitchell/01mitchell.html>

Pacchioli, David. Smart Search. *Research Penn State* 24:2 (2003).

<http://www.rps.psu.edu/0305/search.html>

Sanett, Shelby. The Cost to Preserve Authentic Electronic Records in Perpetuity: Comparing Costs across Cost Models and Cost Frameworks. *RLG DigiNews* 7:4 (2003).

<http://www.rlg.org/preserv/diginews/diginews7-4.html>

Staples, Thornton et al. The Fedora Project: An Open-Source Digital Repository Management System. *D-Lib Magazine* 9:4 (April 2003).

<http://www.dlib.org/dlib/april03/staples/04staples.html>

<http://www.fedora.info/>

Witten, Ian H. et al. Greenstone Open-Source Digital Library Software.

D-Lib Magazine 7:10 (October 2001).

<http://www.dlib.org/dlib/october01/witten/10witten.html>

Book Chapters, Papers , and Proceedings

Belcher, Martin et al. Harvesting, Indexing and Automated Metadata Collection. In *DESIRE Information Gateways Handbook*.

<http://www.desire.org/handbook/3-4.html>

Calado, Pável P. et al. Tools for Building Digital Libraries: The Web-DL Environment for Building Digital Libraries from the Web. *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*. Houston, May 27-31, 2003.

[http://delivery.acm.org/10.1145/830000/827201/p346-](http://delivery.acm.org/10.1145/830000/827201/p346-calado.pdf?key1=827201&key2=1764383601&coll=GUIDE&dl=ACM&CFID=11792774&CF)

[calado.pdf?key1=827201&key2=1764383601&coll=GUIDE&dl=ACM&CFID=11792774&CF](http://delivery.acm.org/10.1145/830000/827201/p346-calado.pdf?key1=827201&key2=1764383601&coll=GUIDE&dl=ACM&CFID=11792774&CF)

[TOKEN=13954292](#)

Crimmins, Francis. *Focused Crawling Review*. 2001. <http://dev.funnelback.com/focused-crawler-review.html>

Heydon, Allan; Najork, Marc. *Mercator: A Scalable, Extensible Web Crawler*. 1999. <http://research.compaq.com/SRC/mercator/papers/www/paper.html>

Liddle, Stephen W. et al. On the Automatic Extraction of Data from the Hidden Web. 2001. <http://www.deg.byu.edu/papers/daswis01.pdf>

Liu, Xiaoming et al. DP9 - An OAI Gateway Service for Web Crawlers. *Proceedings of the Second ACM/IEEE Joint Conference on Digital Libraries*, Portland Ore., July 14-18, 2002. http://www.cs.odu.edu/~liu_x/paper/dp9/dp9.pdf

Lyman, Peter. Archiving the World Wide Web. In *Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving*. Washington DC: Council on Library and Information Resources, Library of Congress, April 2002. pp. 38-51. <http://www.clir.org/pubs/reports/pub106/pub106.pdf>

Lynch, Clifford A. Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central Role of Trust. In *Authenticity in a Digital Environment*. Washington DC: Council on Library and Information Resources, 2001. pp. 32-50. <http://www.clir.org/pubs/reports/pub92/lynch.html>

Palmieri, Juliano Lage et al. Web Services and Performance Evaluation: Collecting Hidden Web Pages for Data Extraction. In *Proceedings of the fourth international workshop on Web information and data management*. McLean, Virginia, USA. November 08, 2002. <http://doi.acm.org/10.1145/584931.584946>

Pierre, John M. Practical Issues for Automated Categorization of Web Sites. *ECDL Workshop on the Semantic Web*. Lisbon, September 21, 2000. <http://www.ics.forth.gr/isl/SemWeb/proceedings/session3-3/paper.pdf>

Raghavan, Sriram and Hector Garcia-Molina. Crawling the Hidden Web. In *Proceedings of the 27th International Conference on Very Large Data Bases (VLDB 2001)*. <http://www-db.stanford.edu/~rsram/pubs/techreps/hiddenWeb.pdf>

Monographs and Reports

Arms, William Y. *Web Preservation Project Interim Report*. Ithaca, NY: Cornell University, January 15, 2001. <http://www.cs.cornell.edu/wya/LC-web/interim.doc>

Christensen-Dalsgaard, Birte et al. *Final Report for The Pilot Project "netarkivet.dk"*. 2003. <http://www.netarkivet.dk/rap/index-en.htm>

Day, Michael. *Collecting and Preserving the World Wide Web: A Feasibility Study Undertaken for the JISC and Wellcome Trust*. UKOLN University of Bath, 2003.

http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf

Nichols, Stephen G. and Abby Smith. *The Evidence in Hand: Report of the Task Force on the Artifact in Library Collections*. Washington DC: Council on Library and Information Resources, 2001.

<http://www.clir.org/pubs/reports/pub103/contents.html>

Preservation Metadata for Digital Objects: A review of the State of the Art. A White Paper by the OCLC/RLG Working Group on Preservation Metadata. January 31, 2001.

http://www.oclc.org/research/projects/pmwg/presmeta_wp.pdf

Trusted Digital Repositories: Attributes and Responsibilities. An RLG-OCLC Report. Mountain View, CA: Research Libraries Group, May 2002.

<http://www.rlg.org/longterm/repositories.pdf>

Web Sites

Autonomy. <http://www.autonomy.com>

CiteSeer Scientific Literature Digital Library. <http://citeseer.nj.nec.com>

Entrieva. *Semiotagger*.

<http://www.entrieva.com/entrieva/products/semiotagger.asp?Hdr=semiotagger>

GammaSite. <http://www.gammasite.com>

HTTrack. *HTTrack Website Copier*. <http://www.httrack.com/>

Interwoven. *Interwoven Metatagger*.

http://www.interwoven.com/products/content_intelligence/index.html

klarity. <http://www.intology.com.au/20products/50Klarity/>

Koster, Martijn. *The Web Robots Pages*. <http://www.robotstxt.org/wc/robots.html>

LOCKSS [*Lots of Copies Keeps Stuff Safe.*] <http://lockss.stanford.edu/>

NetLab. *The Combine Harvesting Robot*. 2000. <http://www.lub.lu.se/combine/>

Search Engine Showdown. <http://www.searchengineshowdown.com>

Search Engine Watch. <http://www.searchenginewatch.com>

SearchTools.com. *Search Tools Listings—Tools for Taxonomies, Browsible Directories, and Classifying Documents into Categories*. <http://www.searchtools.com/info/classifiers-tools.html>

SWISH-E. <http://swish-e.org/>

WebQL. *WebQL: A Software Tool for Web Mining and Unstructured Data Management*. http://www.caesius.com/unstructured_data_tool.html

Unpublished Sources and Documentation

Bibliothèque Nationale de France. *Crawler Information Processing*. (Unpublished notes) 2003.

Internet Archive. *Archive Open Crawler Project: Overview*. (Unpublished notes) 2003.

PANDORA Project. *PANDAS Manual*. 2003. <http://pandora.nla.gov.au/manual/pandas/>

Rissanen, Mika, Juha Hakala and Kaisa Kaunonen. *Manual for Installation and Usage of the NEDLIB Harvester*. 2001.
<http://www.csc.fi/sovellus/nedlib/ver11/documentation11.doc>

Shirky, Clay. *Digital Archiving: Harvesting*. (Unpublished report) 2002.