NDIIPP Web-at-Risk: The Development of a Web Archiving Service at the California Digital Library

Tracy Seneca and Shifra Pride Raffel

California Digital Library, Oakland, CA, USA
http://www.cdlib.org
tracy.seneca@ucop.edu, shifra.raffel@ucop.edu

Abstract. The California Digital Library is building a Web Archiving Service (WAS) to be used by multiple institutions to capture, curate and preserve government publications on the web. This work is undertaken with the University of North Texas, New York University and others as part of the Web-at-Risk NDIIPP grant. The Web-at-Risk project consists of four paths of activity: web archiving needs and usability assessment, partnership building for collaborative web archiving endeavors, experimentation and research into promising web archiving technologies and strategies, and the development of the WAS. The WAS is a Java-based, open-source toolset that integrates existing applications from CDL, as well as third-party software such as Heritrix and JHOVE, and is being released in eight distinct stages. One year into this three-year project, extensive surveys, interviews and focus groups have been conducted for needs assessment, and the Heritrix web crawler has been successfully embedded in the Web Archiving Service application, which stores capture results in CDLs Digital Preservation Repository.

1 Web-at-Risk NDIIPP Grant: Project Background

The Web-at-Risk is a three-year project led by the California Digital Library (CDL), whose goal is to build tools that will enable librarians and archivists to capture, curate and preserve web-based government and political information. The primary focus for collection is US federal, state, and local government information, but will also include web documents on political events, campaigns, and social movements from non-governmental organizations, non-profit, and international government sources and also policy documents, campaign literature, and information surrounding local political movements.

This project is one of eight grants awarded by the National Digital Information Infrastructure Preservation Program (NDIIPP) [1]. The work is undertaken by the CDL and its grant partners, New York University and the University of North Texas, with additional support from Stanford University, the San Diego Supercomputing Center, the Arizona State Library and the Library of Congress.

The California Digital Library is the 11th University of California Library, providing digital library services for the entire University of California community. Other University of California libraries are also involved in this grant, with staff contributing their expertise from campuses in Berkeley, Davis, Los Angeles, San Diego, San Francisco and Santa Barbara.

The project's initial proposal states the importance of the materials to be captured:

> To understand the political landscape of a democratic nation - whether historically or in the midst of grappling with contemporary issues - we must confront a vast quantity of government documentation and literature with which political parties, citizen groups, and a myriad of other organizations try to educate the public and influence government. This literature, a critical element of our nation's heritage, is increasingly, and often exclusively, made available on the World Wide Web. Research, public, and special collection libraries have a crucial role to play in providing enduring access to these essential documents. Indeed, their historic missions are necessarily translated into the responsibility for safeguarding our web-based heritage.
>
> Yet these libraries are hampered in taking up their historic preservation roles because they cannot individually build the considerable technical infrastructure that is required for archiving web-based information [2].

The"at risk" designation refers both to the ephemeral nature of web resources in general, and to the particularly unstable nature of local government and political resources. In 2003, the California Digital Library undertook a broad survey of the United States .gov domain, with assistance from Stanford and funding from the Andrew W. Mellon foundation. The report from that survey, "Web-Based Government Information: Evaluating Solutions for Capture, Curation, and Preservation" highlighted the various challenges and risks posed by web-based government publications. These risks include the inability of the .gov domain to accurately represent the true volume and scope of U.S. federal government output, the volatility and loss of content, the wide diversity of formats and genres, and the increasing opacity of these documents as they become accessible only via dynamic forms [3].

These concerns were echoed by the librarians, archivists and researchers who took part in Web-at-Risk focus groups and interviews.

> "One state agency published its annual county-level statistical report on the web in 1998 for the first time. The next year, the agency replaced the 1998 report with the 1999 report. That has pretty much become our standard bad example." [4]

> "State legislatures dont usually archive their own materials from the Web. They just replace last sessions materials in favor of this sessions. You cant get at committee assignments from 1999 to 2004." [5]

The web is both the means that these agencies use to distribute their documents and, increasingly, the public face of that agency. The need to appear current and up to date often trumps the archival value of older documents, which can be viewed as a detriment to an agencys web site, however valuable they may be to current and future researchers.

Another related project is the Political Communications Web Archiving (PCWA) project, undertaken by Web-at-Risk grant partner New York University, the Internet Archive and others. That projects work precedes the Web-at-Risk project, and the 2004 PCWA report provides an excellent overview of the range of issues confronted in archiving political web sites[1]. That report raises additional issues contributing to the volatility of this body of information: "Production and maintenance of political Web sites are also affected by other factors, some of them less predictable, such as the financial or electoral fortunes of the producing entity, or government suppression of that entity." [6]

To address these concerns, the CDL is building the Web Archiving Service (WAS), a web capture and curation service that is integrated with CDLs existing Digital Preservation Repository (DPR). The development of this service is guided by the projects curatorial partners, a group of 22 government information specialists from several institutions who serve as the pilot users of the WAS, and whose experience and feedback inform the development of the service. The project curators bring a wide range of knowledge and expectations to the project. Some have no prior experience in web archiving, while others have already constructed significant web archives and will be comparing the tools developed by CDL to tools they already know[2] While the projects primary constituents are the libraries of the University of California, New York University and the University of North Texas, the WAS is being developed as an open-source toolset that can be fully or partially deployed by other organizations.

2 Project Paths

While the development of the WAS is the primary goal of this project, it is not the only planned outcome. The WAS is being developed within a larger

[1] Both the "Political Communications Web Archiving" report and the "Web-Based Government Information" report can be consulted for more comprehensive overviews of the threat to political information on the web, including quantitative studies of information loss on the political web.

[2] Curators with established web archiving experience include Valerie Glenn of the UNTs Cyber-Cemetary project [7], Gabriella Gray, creator of the UCLA Online Campaign Literature Archive [8], and Richard Pearce-Moses of the Arizona State Library and author of the Arizona Model of Web Preservation [9].

context of inquiry involving extensive user needs research, experimentation and policy development. The goal is to deliver a toolset, and in the process, to gain a broader understanding of how web archiving will impact collection development in libraries, what the emerging possibilities are, and how to provide broader support for librarians embarking on web archiving efforts. The project is divided into four paths of activity: development, experimental, partnership building and assessment[3].

2.1 The Development Path

The Web Archiving Service is being developed over the course of two years with release stages encompassing different aspects of the application. While details concerning WAS architecture are included later in this article, Fig.1 provides a general overview of the application scope.

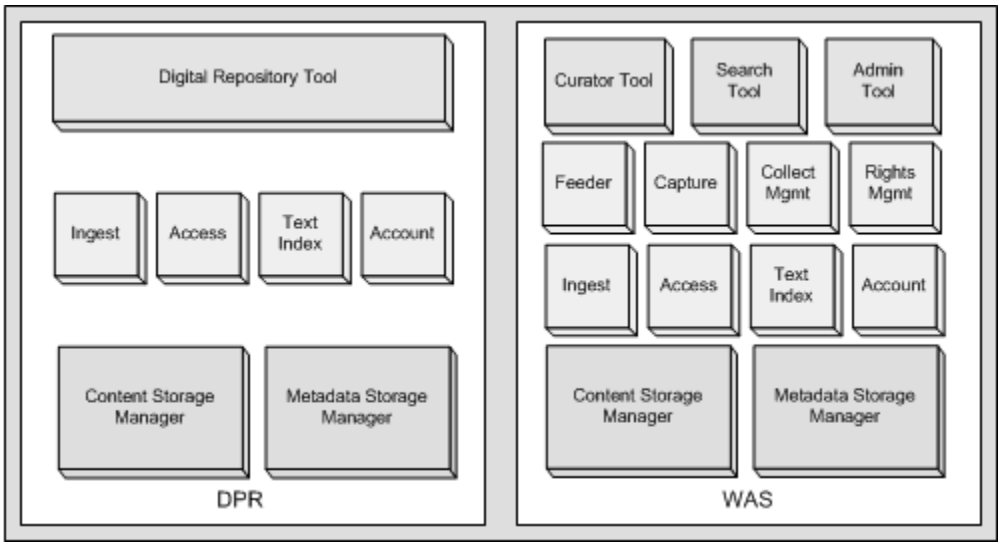

Fig. 1. Components of the DPR and WAS

The Web Archiving Service is being developed using CDLs Common Framework, which is also the technical environment for CDLs Digital Preservation Repository. WAS will rely upon and extend services developed for the DPR. Figure 1 indicates the tools and services already in place for the DPR. Figure 2 represents the tools and services that will be either be created or expanded for

[3] Details concerning the activities, deliverables and timelines for all four paths are available at http://wiki.cdlib.org/WebAtRisk/tiki-index.php?page=ProjectPlans.

the WAS. Key design principles here are the reusability of Common Framework services and the flexibility to integrate open source software, such as the Heritrix crawler developed by the Internet Archive [10], into WAS services and tools.

The WAS is being developed and enhanced over the course of eight release stages, as follows:

Release 1: Basic Capture	July 2006
Release 2: Improved Search and Display	October 2006
Release 3: Improved Analysis and Reports	December 2006
Release 4: Collection Building	February 2007
Release 5: Administration and Rights Management	May 2007
Release 6: Event Capture	July 2007
Release 7: Preservation Features	October 2007
Release 8: Integrate Enhancements	November 2007

Note that at time of this paper (June 2006), the WAS is still in the early stages of development. Consequently, the information that follows concerning specific features, the role of METS and other archival standards, and the details of architecture are subject to change.

Release 1 allows curators to create and run capture jobs using a simplified interface to Heritrix, and ingests capture results, including crawl reports, into CDLs repository. This release relies on the Web Archive Access tool (WERA) [11] developed by the International Internet Preservation Consortium (IIPC) [12] to display captured content; display is not yet integrated with the WAS application.

Releases 2 and 3 will both address the analysis, display and reporting tools that curators require to evaluate the results of a capture. Integrated search tools will be made available in Release 2. Tentative plans are to continue to employ CDL's full text search utility, the eXtensible Text Framework or XTF [13], to index object metadata if not other aspects of the web objects, and to employ NutchWAX [14], Internet Archive's set of web archiving extensions to the Lucene-based search engine Nutch, to index and search the text of the web content. Release 3 will include features to help curators analyze and interpret the results of a web capture.

With release 4, the WAS will enable curators to organize captured materials into discrete collections with the ability to exclude portions of capture results from a collection if needed. Following that, release 5 will include the ability to associate rights metadata with sites either prior to or following capture. This metadata will include the ability to declare the rights status of an item, stating whether the item requires the content owner to be notified of the capture, whether the content owners specific permission is required, or whether the item can be captured without notification or permission. Web-at-Risk staff have been following the ongoing work of the U.S. copyright office with orphan works leg-

islation [15], and took part in the March 2006 Section 108 Study Roundtable in Los Angeles to advocate for library copyright exceptions in web archiving [16].

Release 6 focuses on specialized functionality needed for large event captures in response to sudden and historic events such as Hurricane Katrina. The unique needs associated with different types of web captures have been widely discussed in the literature. Julien Masens [17] points out the distinction between topic-centric and domain-centric crawls in his comparative study of web archiving efforts. Currently all of the Web-at-Risk curators plan to use the WAS to conduct topic-centric crawls, which are distinguished by the curators' expertise in the subject area, and based on selective capture of relevant sites. In event-based captures, however, web sites may be produced quite suddenly on a global scale in such a way that no individual can claim expertise concerning the content.

Approximately six months into the Web-at-Risk project, Hurricane Katrina struck New Orleans, and CDL turned its newly installed crawlers to the task of capturing the ensuing web content. CDL collected over 600 seed URLs from specialists in a broad range of fields covering the social, political and environmental impact of the disaster. This distributed expertise changes the curators' relationship to a collection, and often requires intensively collaborative work across institutions. This can become a challenge if the curator's sole expertise is assumed and built into the data model. Many of the topic experts in the Katrina capture were at Louisiana State University, an organization not affiliated with the project. Event captures such as this benefit from what the Political Communications Web Archiving Report refers to as "federated" archiving, where "important activities, such as selection and storage are distributed, and others, such as data administration and management, are centralized." [6] Release 6 will include tools that allow for site nominations, with the flexibility to respond to ad hoc collaborations and with additional tools to allow a curator to review, accept or reject site nominations.

Release 7 will focus on preservation features and will include routine checksum reports as well as the generation of alternative or desiccated formats of certain file types. This preservation strategy, outlined by John Kunze at the 2005 International Conference on Digital Preservation, is based on the likelihood that the simplest data formats are also the most likely to stand the test of time, and are best produced at the moment of capture [18]. Plain text and JPEG versions of an HTML file might complement the archival copy of that file, in the event that the original should fail to render in the future. A final release is also scheduled to address opportunities for enhancement that may arise over the course of the project.

2.2 The Experimental Path

There is considerably more to investigate and explore in web archiving than a three-year project can possibly integrate into a production environment or end-product. Consequently, the Web-at-Risk project has also outlined an experimental path of activity with three objectives.

First, the Web-at-Risk staff will engage in ongoing research and involvement in the field of web archiving at large. The Web Archiving Service is a relative newcomer to the field, and there are a number of existing points of comparison, such as the PANDAS system used by the PANDORA archive of the National Library of Australia; Archive-It, a system recently released by the Internet Archive; and the Web Curator Tool currently under development at the National Library of New Zealand. Most of the aforementioned institutions are members of the International Internet Preservation Consortium, and have used that venue to share documentation, data models and interface access with other projects. This open attitude in the field has had a direct impact on the WAS architecture. For instance, the nature of the site object in the WAS architecture (detailed in the WAS Data Model section of this paper) was informed in part by the Web Curator Tool data model for targets and groups [19]. This is also a quickly changing field, with new tools becoming available on a regular basis, such as the Heritrix module developed at Emory for conducting adaptive, focused crawling as part of their MetaCombine project [20]. These kinds of tools need to be explored as they become available, with the project remaining flexible enough to respond to them if needed. In addition, Web-at-Risk staff play a role in developing the standards that all of these web archiving projects rely upon. For example, John Kunze of CDL is a contributing author to the forthcoming WARC specification [21], the file format that will replace the current ARC format that most (if not all) of these archives use for storage and exchange [4].

This path also engages in data exchange and replication tests among the various project partners. This work is not expected to become part of the user interface by December 2007, and will not involve the projects curators. However, it does represent a key grant deliverable; all of the NDIIPP-funded projects are required to deliver their archived materials to the Library of Congress at the end of the grant period. It is also a critical point of collaboration between project partners, where staff at New York University and the University of North Texas play important roles.

Finally, as with any software development endeavor, there are a number of features that would be highly desirable for the WAS, but it is unclear whether they can feasibly be included in the production environment. The focused-crawling tools from Emory are an excellent example of this; focused crawling

[4] Because the file format generated by Heritrix crawls is likely to transition from the ARC to the WARC format during the course of this project, we will refer to Heritrix output as the W/ARC format.

would be an excellent component to include with the event capture functionality in Release 6, but it is still too much of an unknown to include in the WAS specifications. There are also confirmed requirements that still need research and testing before they can be implemented. For example, this path will also investigate the automatic generation of derivative or desiccated data formats for selected file types in advance of the preservation features included in Release 7. Thus, the experimental path timeline is composed of a series of investigations and tests, to be carried out in advance of the WAS release to which they pertain.

2.3 The Partnership Building Path

This path is responsible for the business aspects of web archiving. Like many large-scale digital preservation efforts, the Web-at-Risk project includes a complex range of project partners. In addition to the two major project partners, New York University and the University of North Texas, the CDL also serves the individual campus libraries of the University of California, effectively creating several additional project partners. Further, the San Diego Supercomputing Center (SDSC) is providing storage services to CDL, setting up a complex relationship of stewardship between the library capturing the data, CDL, and the SDSC. One task of this path is to draft agreements to clarify the roles and responsibilities of each party in this distributed and collaborative preservation environment. These will include a Model Third-Party Data Stewardship Agreement, a Model Remote Replication Agreement, Web Archiving Service Agreements, and other documents needed for ongoing collaborative web archiving work.

In addition, this path will create a financial sustainability guide, evaluating costs and investigating possible revenue streams to determine how best to carry web archiving projects forward beyond the life of the grant.

2.4 The Assessment Path

The assessment path work is led by Kathleen Murray of the University of North Texas, and includes both initial needs assessment and ongoing usability analysis during the project's development cycles. In the first year of the project this path focused on needs assessment, where activities included an extensive survey of our curatorial partners, initial test captures with subsequent analysis and feedback from curators, and five focus group sessions with librarians and archivists who represent potential users of the service. Additionally, curators conducted in-depth interviews with researchers who represent potential web archive end users, and with government and non-profit agencies representing content owners whose materials would be archived.

Assessment activities continue through the life of the project, as indicated in Fig. 2, becoming closely linked with the release schedule on the development

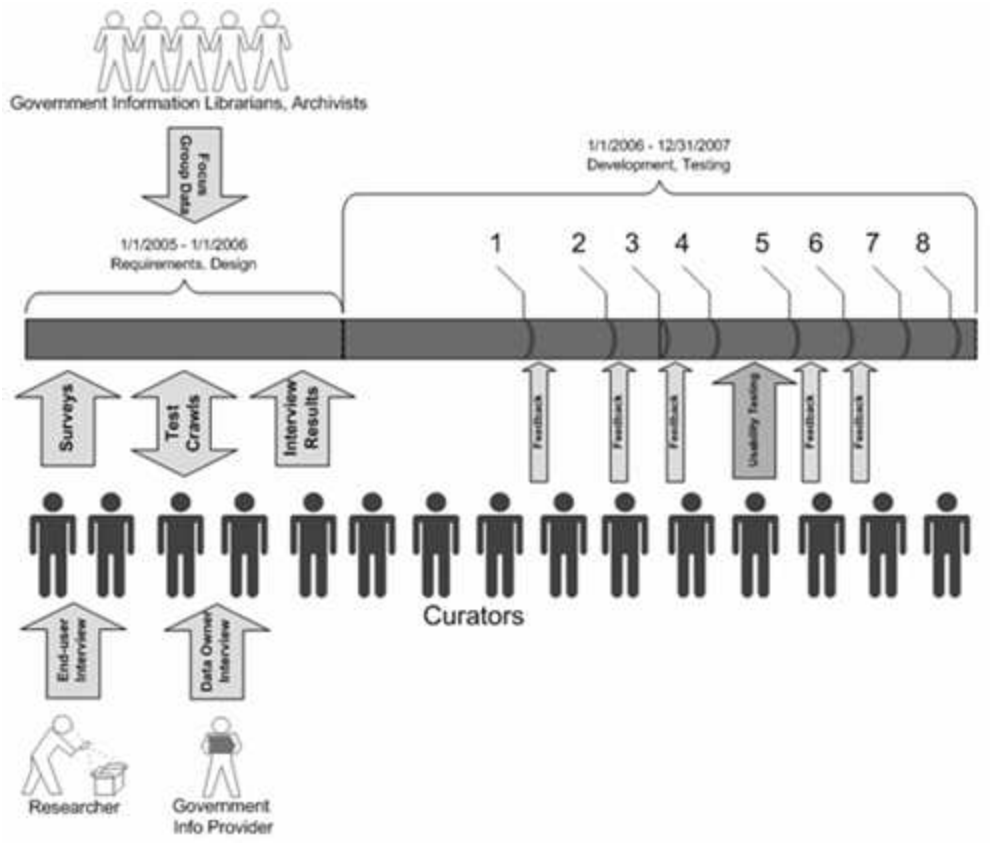

Fig. 2. Assessment activity over the course of the Web-at-Risk project

path. This path will contribute quality-assurance work prior to each WAS release, and will engage in post-release surveys, interviews and formal usability testing.

All of the survey instruments, interview scripts and focus group scripts are available for review or re-use, as are detailed reports from all of these assessment activities [22]. The findings of the needs assessment work are broader than the scope of this paper, but some findings have influenced WAS design and the development path. Here are a couple of examples of how the assessment and development paths interact, and some of the challenges involved.

2.5 Pilot Users and Web Archiving Requirements

An initial needs assessment survey of 58 questions was conducted with the Web-at-Risk curatorial partners. The survey covered issues from selection to preservation and WAS user interface needs. The projects focus on government and political information had an interesting impact on the respondents expectations and assumptions about web archiving. As previously mentioned, our curators are all government information subject experts, as were most of our focus group participants. This lent a considerably more document-centric view to their expectations concerning web archiving. In many cases, the documents they now seek to capture and archive from the web have had a rich history and can seen as a continuation of a series that was previously available in print. In this context, the web site itself is not the object of capture. From the subject experts perspective, the web site may in fact serve as a nuisance, and tools built to archive the entire site effectively may be seen as having the same poor signal-to-noise ratio as the web itself. Only 44% of the respondents reported that they plan to collect websites in their entirety. The remainder plan to collect documents at a more granular level. Only half felt that it was "important for end users to interact with archived materials in a fashion that mirrors the source materials from the time of capture" [23].

These findings pose some interesting problems. While the Web Archiving Service must be built to solve the specific problems encountered by grant participants, it also needs to serve broader purposes beyond the life of the grant. Ultimately, the WAS should effectively capture content on any topic, not just political and government information. In this case, it is a happy accident that many of the document types our curators hope to collect, such as environmental impact reports, tend to be published the PDF format, so a capture filtered to only gather PDF files and perhaps MS Word documents might do a reasonable job of easily capturing a large collection of reports. It is clear that this option will need to be available in the user interface to test this assumption early on. There is unfortunately no reason to assume that this rough correlation between a document type and a file format will last, as new formats and means of publication become available.

2.6 Integrating Assessment

The bulk of Web-at-Risk assessment work has been designed and conducted by project staff at the University of North Texas, with input and assistance from other project partners. This division of labor has allowed for a significant amount of assessment work throughout the project, but poses some challenges as well. Even within an organization, it can sometimes be difficult to interpret and integrate assessment results into design. Here, Web-at-Risk staff need to take extra steps to ensure that assessment results are communicated to the design team and inform their work.

3 Web Archiving Service Architecture

While the four paths of activity outlined above make up the Web-at-Risk project, the heart of the project is the development of the Web Archiving Service. The following sections will provide further detail about the WAS architecture, data model, the technologies used for development and storage, the specifics of the capture service, and the role that METS plays in the overall design.

3.1 The Common Framework

The Web Archiving Service, as mentioned previously, is being developed in the context of the CDL Common Framework, the current technical environment for the Digital Preservation Repository. The services represented earlier in Fig.1 are examples of Common Framework services, which are designed to be modular and reusable. The ingest, access, feeder, text-index and account services will all be reused in the Web Archiving Service. These web services are built in Java (currently version 1.4) using a common base of architecture and extensible classes on both the client and the service side.

Common Framework user interfaces use JSP pages to invoke Struts action classes which in turn call Common Framework clients. Clients use either SOAP [24] or REST [25] to communicate with a Common Framework service. The Capture Service and Capture Agent Service (discussed in detail below) are both examples of Common Framework SOAP services.

Common Framework SOAP services are deployed as .war files in a servlet container (currently Jakarta Tomcat 4.1.29). Each service uses a servlet class (the "Receiver") to exchange SOAP messages with clients. Common Framework SOAP clients and receivers use the SAAJ Java API [26] to produce and interpret SOAP. Other languages and APIs could be used for creating and parsing SOAP, as long as they support attachments. Common Framework SOAP clients and services must use a common message API expressed in SOAP, but have no other dependencies on each other.

Each SOAP message comes to the receiver bearing a request type and other information. The receiver delegates each request type to a different "handler" class. New request types and handlers can be added without any change to existing code, using properties files. This allows tremendous flexibility and extensibility.

The Common Framework also employs CDL-developed plug-ins, such as the Nice Opaque Identifier (NOID) [27] for persistent Archival Resource Key (ARK) [28] generation, and the eXtensible Text Framework (XTF)currently used for text indexing, searching and browsing. Third-party software used by the Common Framework includes JHOVE [29], the Storage Resource Broker (SRB) [30],

and MySQL.

3.2 The WAS Data Model

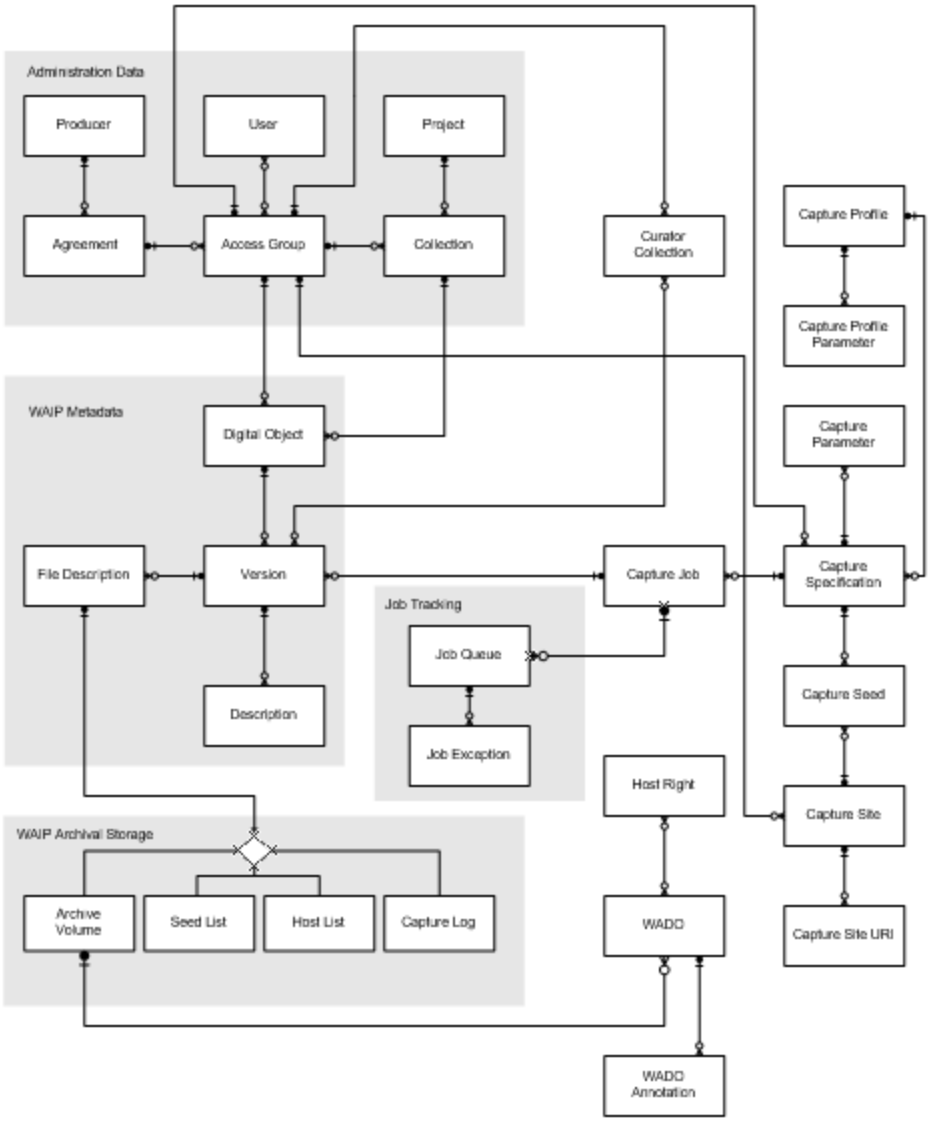

Fig. 3. Web Archiving Service data model

Some basic elements of the Web Archiving Service data model (Fig. 3) are the following:

Site: A set of one or more URLs that defines an entity to be captured. Additional metadata may be associated with a site.
Capture Specification: A set of parameters that define the desired behavior of a capture. This includes the specific sites designated for capture.
Capture Profile: A set of default values for capture parameters that simplifies the creation of Capture Specifications.
Capture Job: A run of a capture of one or more Sites using a defined Capture Specification.

Two elements of this data model are particularly important to the WAS approach to web archiving; these are the Web Archive Information Package (WAIP) and the Web Archive Digital Object (WADO). The WAIP, illustrated in Fig. 4 is the aggregate of all information resulting from the execution of a Capture Job. This includes captured content, metadata, capture logs and capture reports.

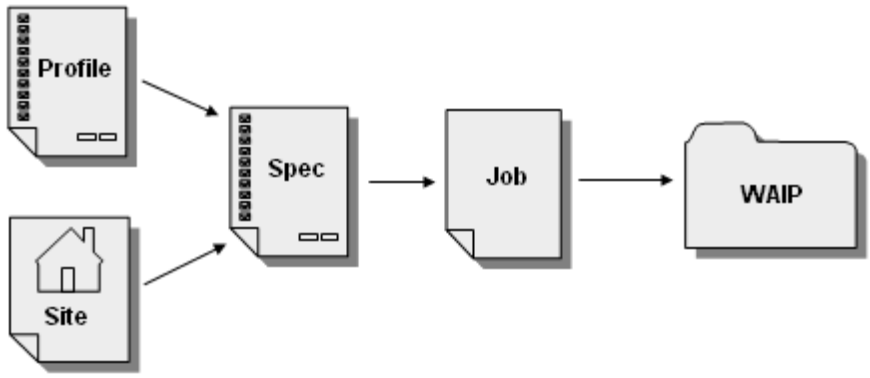

Fig. 4. Capture data flow and the WAIP

The WAIP represents an entire capture; the WADO, conversely, corresponds to an individual item resulting from a Capture Job. The WADO may be at the level of a single image embedded in a web page and will include any associated metadata, such as rights information pertaining uniquely to that object. The distinction between the WAIP and WADO was made in an effort to separate the complex and sometimes conflicting preservation and access needs within the web archive. This complexity is particularly challenging in an environment where captured content might be shared across multiple collections owned by different

curators. Fig.5 illustrates how the WAIP and WADO differ and interact.

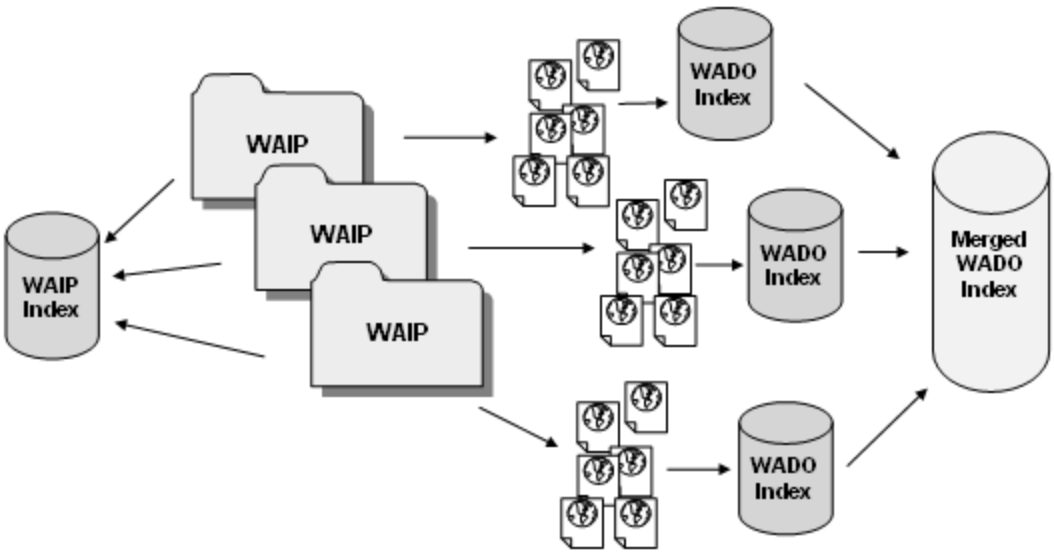

Fig. 5. Indexing captured content

For example, to ensure the authenticity of captured content, the WAIP will be stored exactly as originally captured. However, curators should also be able to benefit from and build upon previously captured web materials in new contexts. WAS designers initially explored a model where curators could import or share crawls across multiple collections. However, the likelihood is high that entire capture jobs, particularly those with many seed URLs, will contain a great deal of information that is not relevant to multiple collections. Sharing content at the level of WAIPs, or entire capture results, creates an all-or-nothing requirement for relevance. The WADO index, by contrast, allows a finer granularity for sharing captured results across collections. If a curator captures a particular page, and older versions of that page have already been archived by someone else, the archive should be capable of linking the curator to the relevant archived pages. Figure 8 further illustrates the distinction between the archival context of the WAIP and the access/display context of the WADO.

3.3 The Capture Service

The WAS has necessitated development of a new Common Framework web service: the Capture Service. This service communicates directly with the UI and

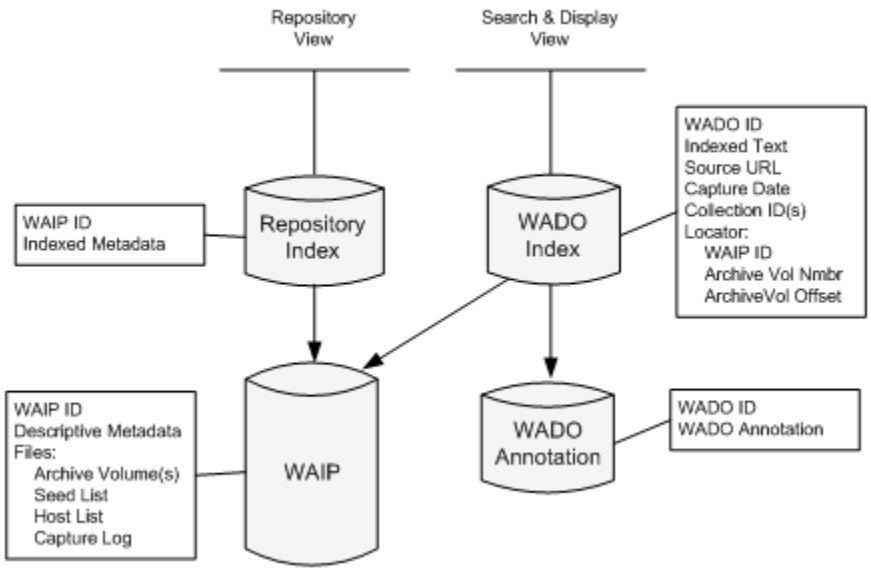

Fig. 6. The WAIP and the WADO

performs most functions of the WAS. Another web service layer, the Capture Agent Service, is contained within the Capture Service. The Capture Agent Service wraps individual instances of software (e.g., Heritrix) to communicate with the Capture Service.

The Capture Service takes all requests for capture execution, and for the creation and editing of capture specifications and profiles. The Capture Service then queries a database to find appropriate capture agent instances (which can be local or remote), and relays capture execution requests to them. CDL has used the term "capture" rather than "webcrawl" with the idea that the WAS might be made extensible for other varieties of harvest (OAI-PMH, etc). Such flexibility is only theoretical at this point. As a first step towards non-webcrawl extensibility, the capture service has been protected from the detailed workings of the harvester(s) it controls. Given the modularity of the Common Framework architecture, another implementation of this service should require only extension of base classes and addition of properties.

To keep the Capture Agent Service agnostic about the type of software it wraps, a"capture controller" class has been introduced. As with all Common Framework services, the Capture Agent Service forwards each request type to a different handler class. In the Capture Agent Service, each handler then invokes the capture controller specific to the variety of harvester and to the request type. This capture controller contains the code that instantiates and communi-

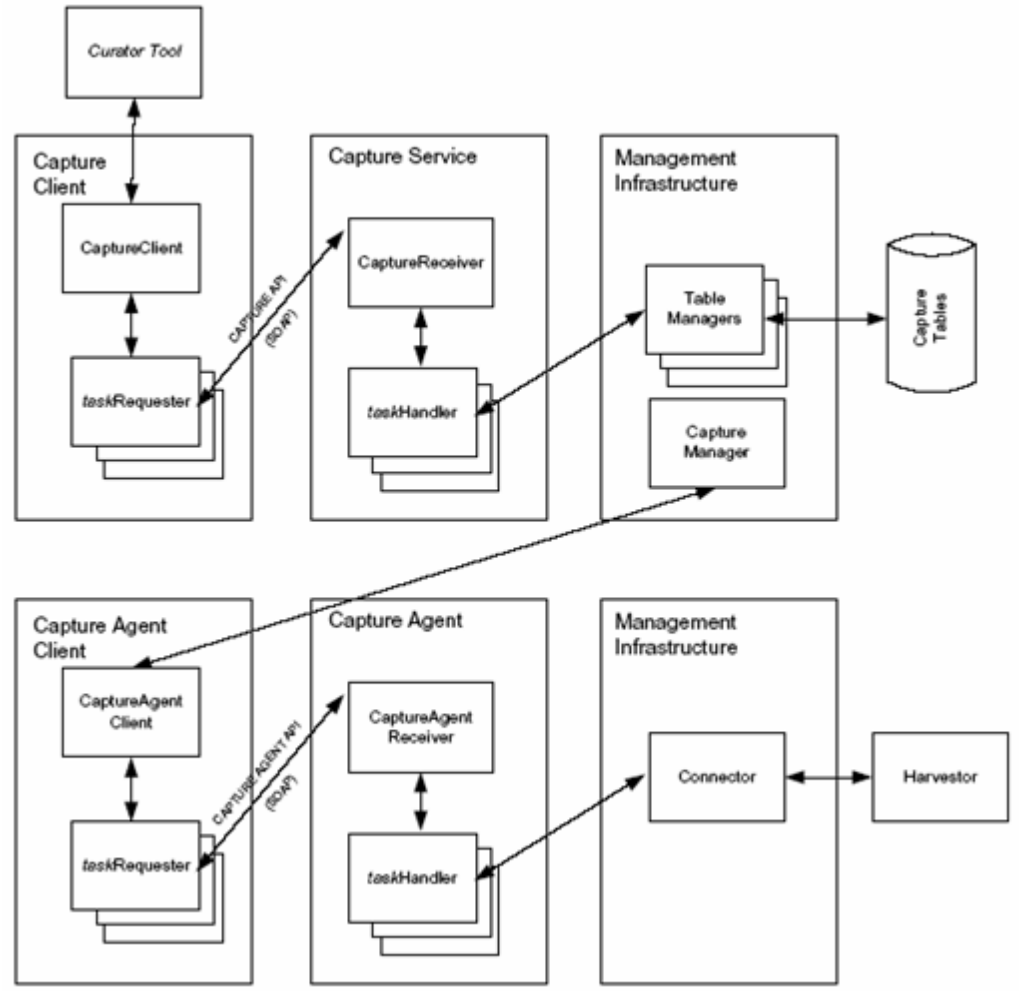

Fig. 7. The Capture Service and the Capture Agent

cates with a single Heritrix instance. In the future, the Capture Agent service could also wrap non-webcrawl capture software with the simple addition of new properties and capture controller classes.

XML-formatted parameters for the capture are sent to the Capture Service (and thence to the Capture Agent Service) in an attachment to a SOAP message, and the parameters are transformed via XSLT into the order.xml document required by Heritrix for each crawl.

The "createCaptureJob" capture controller invokes NutchWAX to generate an index for the W/ARC products of the crawl. This capture controller then calls a WAS-specific feeder service to generate a METS file and deposit the WAIP in the Digital Preservation Repository.

3.4 Embedding Heritrix

As seen above, CDL has chosen to construct a Capture Agent wrapper rather than to follow the Internet Archive approach to controlling multiple crawler instances. The following section of this paper discusses the technical details, advantages and challenges of embedding Heritrix in this way.

Embedding is not the only strategy. To call the Heritrix crawler from a Java application, it is not necessary to locally instantiate the org.archive.crawler.Heritrix class. If the calling application is being run under Java 5, and if a port local to the Heritrix instance is available to serve as the JMX port, the Heritrix binary can be contacted and controlled using JMX, the Java 5 remote monitoring and management technology [31]. Clients that exploit the exposed JMX operations in Heritrix include the Internet Archive's JMX client [32], the JConsole tool included with the Java 5 SDK [33], and the Internet Archive's Heritrix Cluster Controller (HCC) for managing multi-machine groups of Heritrix crawlers [34].

The Internet Archive uses their JMX client and HCC together to enable scaling to multi-machine crawling. Although scalabity is crucial for the WAS, several factors spurred the development of the Capture Agent Service rather than the adoption of the Internet Archive's strategy. It is advantageous for the Capture Service to register instances of capture software like Heritrix, and their capture agent wrappers, in a persistent directory or database so as to track their individual status systematically. HCC, rather than registering instances of Heritrix persistently, does a lookup in the directory service to which the jndi.properties points, and dynamically discovers whatever Heritrix instances are registered with that service. Although this dynamic lookup allows for more spontaneous addition and removal of crawler instances, it does not necessarily preserve state information about the lifecycle of any given instance.

In addition, CDL's Common Framework services currently run under Java 1.4, which does not support JMX. Thus it was necessary to develop the Capture Agent layer for the Capture Service, instead of employing the Internet Archive's existing approach. Both Heritrix and HCC, the Internet Archive's crawler-controller service, will be embedded in the Capture Agent, but HCC will be used more for monitoring than for crawler control. An advantage to the JMX approach is that the invoking application does not necessarily have to reside on the same machine as the Heritrix instance. By contrast, the Capture Agent Service is designed to operate solely on a local, embedded instance of Heritrix, rather than an instance on a remote machine. However, since the Capture

Service can make remote calls to the capture agent instances, the disadvantage of this approach is mitigated.

A now-deprecated approach to embedding the Heritrix crawler in a Java application was to instantiate a CrawlJobHandler [35]. Then the Internet Archive added Heritrix functionality for bundling the crawler as a web application, instantiating it from a WebappLifecycle class [36] which receives server lifecycle notifications[5]. Following the example of the WebappLifecycle, the Capture Agent Service allocates a new Heritrix object with a single "true" argument to the constructor (meaning true, register this Heritrix with JMX), and ends the Heritix lifecycle by calling the destroy() method.

Some issues worth noting arose when embedding Heritrix in the Capture Agent Service. JMX registration happens automatically in the presence of a JDK (not just a JVM) v. 1.5.0 or greater. When JMX is not enabled for a Heritrix instance, whether because of Java version issues or for lack of the com.sun.management.jmxremote system variable, the registration failure is silent. Null pointer exceptions will be thrown if reference is made to the nonexistent MBeanServer, and the JConsole tool will fail to connect to the specified JMX port (since no MBeanServer is available there), but no informative error is thrown, which can make troubleshooting burdensome. For successful JMX registration to occur, the following either need to be set as system properties or passed as -D command line options:

```
com.sun.management.jmxremote=true,
com.sun.management.jmxremote.port=available port
```

Additional options allowing password-free authentication may be set in a development environment. A jndi.properties file must also exist so that Heritrix can register with the directory service to which the jndi.properties file points; again, if that file is missing, the registration failure is understandably silent.

Customizing the Tomcat instance for Heritrix was non-trivial. The Heritrix logging configuration may need to be migrated to that of the Capture Agent Service, since the current logging level setting impacts the entire servlet container. Classpath issues are also crucial: because the Internet Archive has added functionality to packages such as apache.commons.httpclient, the Heritrix classes must be at the front of the Java classpath. Otherwise Heritrix fails in an uninformative manner, seeming to crawl but without retrieving any content. Although these current issues have obvious drawbacks when including Heritrix as part of

[5] The WebappLifecycle class implements javax.servlet.ServletContextListener, so that it can receive notification from servlet containers (like Tomcat and JBOSS) of application startup and shutdown events via its contextInitialized() and contextDestroyed() methods. When the servlet container lifecycle events are triggered, the WebappLifecycle listener accordingly creates or destroys a crawler instance.

a web service, Internet Archive has been quite responsive, and future Heritrix modifications may include helpful configuration methods such as retrieving options from the web application's web.xml configuration file. In the interim, the Web-at-Risk project has chosen to set Java options and classpath in the shell scripts that start the Tomcat instance, making automated deployment somewhat cumbersome.

Future work on the Capture Agent Service will involve accessing information about the currently running Heritrix job using methods in the CrawlJob #invoke method [37]. Other developments may also include making the built-in Heritrix Jetty web UI available as an alternative to JMX for non-technical staff, or those without access to the Java 5 JConsole tool.

3.5 The Role of METS

Metadata for web content is a subject obviously still replete with open questions and in need of exploration. This exploration is occurring on a number of fronts, and the Web-at-Risk project is following and contributing to the work of a number of different groups concerned with web-archiving metadata. The IIPC has proposed a web-archiving metadata set, and their work was summarized by Julien Masans at the 2005 International Web Archiving Workshop [38]. This metadata set is focused largely on specifying what data needs to be recorded at what level of granularity, differentiating, for example, between metadata that applies to the document from metadata that applies to the overall crawl. The Research Libraries Group (RLG) is also forming a working group on Descriptive Metadata for Web Archives, in which Web-at-Risk staff will be taking part [39]. It is anticipated that this work will also focus on specifying what metadata is recorded. Another related development is the publication of the Preservation Metadata Implementation Strategies (PREMIS) report, which includes an extensive data dictionary specifying preservation metadata for objects, events, rights and other entities [40]. RLG was a sponsor of this work, and Web-at-Risk staff were on the PREMIS working group, so it is expected that the efforts of the PREMIS working group will serve as a source of input. Finally, the question of web archiving metadata naturally arises in any data model for existing web archiving systems, and as mentioned earlier, these alternative data models are being carefully explored by Web-at-Risk staff.

In addition to the question of which metadata is required, there is also the question of how that metadata will be stored in the system architecture. METS, the Metadata Encoding Transmission Standard [41], now widely used by a number of preservation and digital library systems, is one obvious route to explore. The METS XML standard allows for descriptive, administrative, and structural metadata to be associated with digital objects, and can potentially be used to serve this role in web archiving. The question of how to best record web archiving

metadata using METS has been previously explored in some detail by Web-at-Risk project partners at New York University, who outlined an approach to using the METS structMap to associate website components [42]. In this case, METS files can describe components of a web page or components of a web site. This approach was used in the Political Communications Web Archiving project, mentioned earlier. A critical aspect of this approach is that the METS file is not merely descriptive, but is the basis for a METS viewer which "re-assembled, presented, and allowed viewing and manipulation of archived sites" [6].

3.6 The ECHO Depository METS Approach

The role of METS in web archiving is also being explored by the ECHO Depository NDIIPP grant [43]. This group is currently developing the ECHO DEPository METS Profile for Archival Digital Repository Interoperability. This profile has not been officially released, but is designed to package both web archiving products and other objects in a neutral form that allows exchange among digital repositories. The profile working group is headed by Thomas Habing of UIUC.

The descriptive metadata section of ECHO DEPository profile requires MODS descriptive metadata. MODS, a relatively new metadata standard, carries MARC fields into an XML schema and provides more richness of expression than the simpler Dublin Core metadata standard [44].

3.7 The WAS METS Approach

The CDL Digital Preservation Repository currently requires METS in order to ingest objects into the repository, and since the WAS is built upon the existing DPR, METS will play a role here as well. Whether the Web Archiving Service will adopt the ECHO DEPository profile is uncertain, in part because the level of object granularity differs markedly between the two projects. The ECHO DEPository project plans to decompose the W/ARC files produced by Heritrix into individual files at the level of each HTTP get command (such as an image embedded in a web page), and to represent each of those files in the structMap and fileSec portions of the METS file. By contrast, the Web-at-Risk project plans to apply METS at the WAIP level, discussed earlier. In other words, WAS METS files will describe the entire product of the crawl, including W/ARC files, logs, probably indexes to the W/ARC files generated by NutchWAX, and even the order.xml file which holds configuration parameters. Curators will have the opportunity to attach metadata to the entire crawl, which could include files from multiple domains and websites.

The fileSec and structMap sections would include references to the individual W/ARC files that constitute the WAIP, but not each individual file/product of an HTTP request contained within the W/ARC files. This decision has been

taken almost completely from a pragmatic standpoint, in order to facilitate scalability. It remains to be seen whether this project's approach to object granularity precludes usage of the ECHO DEPository METS profile.

Note that the WAS' more coarsely-grained approach to METS does not interfere with the ability to record metadata elsewhere in the WAS architecture and to link that metadata to individual captured objects. It is clear that Web-at-Risk curators want the ability to provide metadata at the object, page, site and capture levels, which the WAS will offer. METS will not be the means for doing so, nor the means for displaying archived content.

Instead, additional metadata will likely be stored in database tables. However, none of the details of metadata storage - of METS profile usage or of MODS implementation - need be visible to curators. Rather, curators will supply metadata in web forms (with descriptive field names such as "title" or 'description") from the WAS user interface. The WAS will transform this metadata into the appropriate format. Thus curators will not be required to have expertise with any particular standard in order to use the service. The critical principle behind the WAS approach is that the implementation of METS should not pose a barrier to web capture.

3.8 Storage

The Common Framework's storage layer is designed to be modular so that storage technologies can be interchanged with ease. The current Common Framework storage software is the San Diego Supercomputer's Storage Resource Broker, or SRB. SRB has capabilities beyond the Common Framework's needs. For example, metadata can be stored natively alongside the object in SRB using its MetaData Catalog, or MCAT, but the Common Framework stores objects in SRB and their associated metadata in a MySQL database. One SRB feature of which the Web-at-Risk will take full advantage, however, is the ability to replicate stored data automatically from storage site to storage site, with the same object identifier for all copies. Several native replication methods are available, and copies of objects made outside SRB can also be registered within SRB after the fact as replicated copies. CDL and Web-at-Risk partners UNT and NYU plan to test SRB replication with crawled data, and will report the results.

4 Next Steps

At the time of publication, the Web-at-Risk grant is approximately halfway through its three-year grant life, with an aggressive release schedule for the remainder of WAS releases. Follow-up reports detailing each release, as well as

ongoing experimentation with storage solutions and other promising web archiving technologies, will be posted on the project's wiki site [45]. Immediate next steps will be to develop the display and analysis features slated for WAS Releases 2 and 3. This will involve a careful review of the user feedback gathered after the first WAS release and integrating any relevant experimental path work on prototypes and third-party analysis tools. CDL will continue to work closely with the Internet Archive throughout the project to insure that the strategy of embedding Heritrix is effective and successful. Ultimately, as mentioned, CDL plans to extend the Web Archiving Service beyond the focus of political and government information in order to meet the broader and varied needs of the University of California, and to continue exploring effective preservation and storage strategies for the digital environment.

References

1. National Digital Information Infrastructure Preservation Program. Library of Congress. http://www.digitalpreservation.gov/
2. Greenstein, Daniel: The Web at Risk: A Distributed Approach to Preserving our Nations Political Cultural Heritage. (Grant application). 2003. http://wiki.cdlib.org/WebAtRisk/tiki-download_file.php?fileId=143?
3. Web-Based Government Information: Evaluating Solutions for Capture, Curation, and Preservation: An Andrew W. Mellon Funded Initiative of the California Digital Library. California Digital Library. November 2003. http://www.cdlib.org/programs/Web-based_archiving_mellon_Final.pdf
4. Kathleen R. Murray and Inga Hsieh: Focus Group Report: Federal Depository Library Conference - Washington DC - October 2005. March 28, 2006. http://web2.unt.edu/webatrisk/na_toolkit/Reports/fdlc_oct2005_fg_summary_final_28mar2006.pdf
5. Kathleen R. Murray and Inga Hsieh: Web-at-Risk End User Interviews: Summary Report. April 17, 2006. http://web2.unt.edu/webatrisk/na_toolkit/Reports/eu_interview_summary_final_17apr2006_2.pdf
6. Political Communications Web Archiving: An Investigation Funded by the Andrew W. Mellon Foundation. Center for Research Libraries; Latin American Network Information Center, University of Texas at Austin; New York University; Cornell University; Stanford University, Internet Archive. June 2004. http://www.crl.edu/PDF/PCWAFinalReport.pdf
7. CyberCemetary. University of North Texas Libraries. http://govinfo.library.unt.edu/
8. Online Campaign Literature Archive. UCLA Library. http://digital.library.ucla.edu/campaign/
9. Pearce-Moses, Richard: An Arizona Model for Preservation and Access of Web Documents. DttP: A Quarterly Journal of Government Information Practice & Perspective; Spring2005, Vol. 33 Issue 1, p17-24. http://www.lib.az.us/diggovt/GODORT/AzModel_GODORT.pdf
10. Heritrix Home Page. Internet Archive. http://crawler.archive.org/
11. WERA Home Page. International Internet Preservation Consortium. http://archive-access.sourceforge.net/projects/wera/
12. NetPreserve.org: International Internet Preservation Consortium. http://netpreserve.org/about/index.php

13. eXtensible Text Framework (XTF). California Digital Library. March 1 2006. http://www.cdlib.org/inside/projects/xtf/
14. NutchWAX. SourceForge, International Internet Preservation Consortium, Internet Archive, Nordic Web Archive. http://archive-access.sourceforge.net/projects/nutch/
15. Orphan Works. United States Copyright Office. http://www.copyright.gov/orphan/
16. March 8 Public Roundtable Transcripts: Topic 4: New Website Preservation Exception. Library of Congress Section 108 Study Group. http://www.loc.gov/section108/docs/0308-topic4.pdf
17. Masans, Julien: Web Archiving Methods and Approaches: A Comparative Study. Library Trends, Summer 2005, Vol. 54 Issue 1, p72-90.
18. Kunze, John: Future-Proofing the Web: What We Can Do Today. International Conference on Preservation of Digital Objects. September 2005. Gttingen, Germany. http://rdd.sub.uni-goettingen.de/conferences/ipres05/download/Future-Proofing%20The%20Web%20What%20We%20Can%20Do%20Today%20-%20John%20Kunze.pdf
19. Web Curator Tool Project. Software Requirements Specification. National Library of New Zealand. December, 2005. p. 76.
20. MetaCombine Software, Focused Crawling Module. Emory University. http://metacombine.org/software/
21. Information and documentation The WARC File Format. International Organization for Standardization. 2006. http://www.niso.org/international/SC4/N595.pdf
22. Web-at-Risk Assessment Reports. University of North Texas. http://web2.unt.edu/webatrisk/delivs.php
23. Kathleen R. Murray and Inga Hsieh: Web-at-Risk Needs Assessment Survey Report (Abbreviated Version). January 2006. http://web2.unt.edu/webatrisk/na_toolkit/Reports/survey_data_analysis_final_05Jan2006_abbreviated.pdf
24. SOAP Version 1.2. W3C Recommendation 24 June 2003. http://www.w3.org/TR/soap12-part1/
25. Fielding, Roy Thomas: Architectural Styles and the Design of Network-based Software Architectures. Chapter 5: Representational State Transfer (REST). 2000. http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm
26. SOAP with Attachments API for Java (SAAJ). Sun Developer Network. http://java.sun.com/webservices/saaj/index.jsp
27. NOID (Nice Opaque Identifier) Minting and Binding Tool. California Digital Library. http://www.cdlib.org/inside/diglib/ark/noid.pdf
28. Archival Resource Key (ARK). California Digital Library. http://www.cdlib.org/inside/diglib/ark/
29. JHOVE - JSTOR/Harvard Object Validation Environment. ¡ http://hul.harvard.edu/jhove/
30. SRB Main Page. http://www.sdsc.edu/srb/
31. Monitoring and Management Using JMX. Sun Developer Network. http://java.sun.com/j2se/1.5.0/docs/guide/management/agent.html
32. Command-line JMX Client. Internet Archive. http://crawler.archive.org/cmdline-jmxclient/
33. Using JConsole to Monitor Applications. Sun Developer Network. Mandy Chung, December 2004. http://java.sun.com/developer/technicalArticles/J2SE/jconsole.html

34. Overview Heritrix Control Cluster. hcc 0.2.0-200606011937. http://crawler.archive.org/hcc/apidocs/overview-summary.html
35. Stack, Michael: Re: [archive-crawler] Embedding Heritrix. http://groups.yahoo.com/group/archive-crawler/message/2703
36. WebappLifecycle. Internet Archive. http://crawler.archive.org/xref/org/archive/crawler/WebappLifecycle.html
37. Heritrix.java,v 1.134. 2006/06/10. Line 1788. Internet Archive. http://crawler.archive.org/xref/org/archive/crawler/Heritrix.html#1788
38. Masans, Julien: Metadata for Web Archiving. International Web Archiving Workshop, 2005. Vienna, Austria. http://www.iwaw.net/05/masanes2.pdf
39. Research Libraries Group. Web Archiving Program. http://www.rlg.org/en/page.php?Page_ID=399&projGo.x=25&projGo.y=15
40. Library of Congress. PREMIS Home. http://www.loc.gov/standards/premis/
41. Library of Congress. Metadata Encoding and Transmission Standard. http://www.loc.gov/standards/mets/
42. Myrick, Leslie: DSpace and Web Material: Inroads and Challenges. DLF Spring Forum 2005, San Diego, California. http://www.diglib.org/forums/Spring2005/presentations/myrick0504_files/frame.htm
43. The ECHO DEPository Project. University of Illinois, Urbana-Champaign. http://www.ndiipp.uiuc.edu/
44. Metadata Object Description Schema. Library of Congress. http://www.loc.gov/standards/mods/
45. Web-at-Risk Wiki. California Digital Library. http://wiki.cdlib.org/WebAtRisk/tiki-index.php