The Web Archives Workbench (WAW) Tool Suite: Taking an Archival Approach to the Preservation of Web Content

Patricia Hswe, Joanne Kaczmarek, Leah Houser, and Janet Eke

Abstract

The ECHO DEPository (also known as ECHO DEP, an abbreviation for Exploring Collaborations to Harvest Objects in a Digital Environment for Preservation) is an NDIIPP-partner project led by the University of Illinois at Urbana-Champaign in collaboration with OCLC and a consortium of partners, including five state libraries and archives. A core deliverable of the project's first phase was OCLC's development of the Web Archives Workbench (WAW), an open-source suite of Web archiving tools for identifying, describing, and harvesting Web-based content for ingestion into an external digital repository. Released in October 2007, the suite is designed to bridge the gap between manual selection and automated capture based on the "Arizona Model," which applies a traditional aggregate-based archival approach to Web archiving. Aggregate-based archiving refers to archiving items by group or in series, rather than individually. Core functionality of the suite includes the ability to identify Web content of potential interest through crawls of "seed" URLs and the domains they link to; tools for creating and managing metadata for association with harvested objects; website structural analysis and visualization to aid human content selection decisions; and packaging using a PREMIS-based METS profile developed by the ECHO DEPository to support easier ingestion into multiple repositories. This article provides background on the Arizona Model; an overview of how the tools work and their technical implementation; and a brief summary of user feedback from testing and implementing the tools.

LIBRARY TRENDS, Vol. 57, No. 3, Winter 2009 ("The Library of Congress National Digital Information Infrastructure and Preservation Program," edited by Patricia Cruse and Beth Sandore), pp. 442–460

The Web Archiving Problem

The Ubiquitous Web

For a broad range of organizations, websites are now the delivery mechanism of choice for nearly any type of information content. Much of this content is created and disseminated in electronic formats only, with printed copies considered just a courtesy or convenience. The electronic format environment, while expedient for current access purposes, presents challenges for anyone charged with preserving information over time. These challenges include the sheer volume of Web-published information, traditional issues of selection and description, as well the technical challenges associated with long-term preservation of digital objects.

The Challenges of Web Archiving

Volume and Selection of Web Content An immediate challenge of Web archiving is assuring that all content of long-term relevance delivered through the Web is identified and collected (i.e., harvested). Difficulties arise first from the task of selecting pertinent content for preservation from the enormous volume of information streaming from Web servers at any given point in time. Selection decisions will be influenced by the charge of the individual responsible for capturing specific content types (usually a librarian or archivist) based on appraisal or collection development; on policies created in concert with the mission of the institution or organization; and on the audience or user community being served. The sheer volume of content published on the Web makes a fully manual perusal of online resources infeasible. The sheer volume of web-published information is still a major barrier to collecting content.

The Nature of the Web The dynamic nature of the Web also creates problems for selection and harvesting of content. URLs can change overnight; resources can be taken offline with little or no notice; and new, related content can be added in new or different directories than those visited previously by a Web crawler harvesting an organization's website. Although Web crawling automates archiving of a website, it is quite possible for Web crawlers simply to miss content because of a "robots exclusion protocol" (activated by the website's administrator to make parts of a site "uncrawlable") or because of the impenetrable character of the Deep Web (where content, such as a results page to a Web form, is inaccessible to a Web crawler or Web spider). In addition, the vast measure of the Web renders scalable Web crawling an almost intractable technical challenge. Knowing where to find all content eligible for harvesting according to collection development and appraisal policies becomes nearly impossible without intentional coordination or without Web crawling tools and resources that are designed for, and take account of, the fluid nature of website content and the massive scale of the Web.

The Importance of Context Context is about understanding relationships between different and discrete pieces of information. It is about understanding why the information was created, by which individual or organization, and at what point in time. Contextual information can help define the boundaries and the scope of harvested content.

As with analog objects, much of the usefulness of digital objects, which make up our cultural record, depends on having descriptive and contextual information about them. Once content is identified and harvested, it is necessary to provide access to the digital object. Such content access means that attention should be paid to capturing accurate metadata along with the content itself. This contextual metadata will help describe the origin or "provenance of the resource," as well as why and when it was created. For example, is the discovered resource one in a series of annual reports from a particular state agency? Is it a single publication summarizing research findings? Or does it encompass results from a specific survey taken as part of a larger effort to revamp community services? In the case of a digital object, metadata not only supports human interpretation of content, it is needed to provide crucial technical information for maintaining long-term viability of the object itself.

An Archival Approach to Web Archiving (the Arizona Model)

Foundational Elements of the Web Archives Workbench

The Web Archives Workbench tool suite is based on the principles of the "Arizona Model," an aggregate-based approach to Web archiving designed to bridge the gap between human selection and automated capture. "Aggregate-based" means that rather than archive items singly, or individually, they are organized (grouped) in series, or in aggregates. The Arizona Model was developed in 2003 by Richard Pearce-Moses of the Arizona State Library and Archives.

Background on the Arizona Model

Most state libraries and archives have mandates to collect state agency publications and make them available to the public. To this end, there are well-established depository systems that have worked with paper publications for many years. In a Web environment the nuances of determining what a publication is, or who is responsible for selection and collection of particular information resources, becomes less clear. Nonetheless, to meet these mandates librarians and archivists still must identify, select, acquire, describe, and provide access to state agency information "published" on websites.

In early attempts to develop a collection of state agency electronic publications, two approaches came about. According to Cobb, Pearce-Moses, and Surface (2005), the first approach has its premise in "traditional library processes of selecting documents one by one, identifying

appropriate documents for acquisition; electronically downloading the document to a server or printing it to paper; then cataloging, processing, and distributing it like any other paper publication" (175). While this approach ensures that valuable documents will be gathered, its dependence on manual selection will limit archiving to only a very few items. Scaling this process in accordance with the vastness of Web-based documents would necessitate an expansion in personnel that few state libraries have the funding to address (Cobb et al., 2005). Alternatively, in the other approach, software tools that automate regularly occurring Web crawls are engaged. As Cobb, Pearce-Moses, and Surface (2005) assert, this model "trades human selection of significant documents for the hope that full-text indexing and search engines will be able to find documents of lasting value among the clutter of other, ephemeral Web content captured in the process" (176). Yet, while this model relieves librarians and archivists of the upfront onus of selection and organization, at the same time it may unduly burden future searchers, if full-text indexing and search capabilities do not evolve as anticipated.

The Arizona Model, explained in detail below, constitutes a third approach to Web archiving, incorporating both human assessment and automated tools.

An Archival Approach

The Arizona Model applies an archival perspective to curating collections of Web publications. It exploits certain telling parallels between websites and archives: namely, the concept of provenance (i.e., documents classed together stem from the same source) and the organizational structure inherent in both these kinds of collections—directories and subdirectories for websites, and series and subseries for archives (Cobb et al., 2005). In theory, if websites organize Web publications using common file directory structures, information about individual documents within sub-directories could be inherited from parent directories.

In the Arizona Model, which draws on basic archival practice, websites are handled as hierarchical aggregates rather than as individual items, and the original order of the documents (the order in which the creating agency oversaw them) is maintained. Provenance and original order are considered important contextual pieces of information. Retaining documents in the order in which they were originally managed and keeping them clustered together based on the originating agency enhance one's knowledge of the creation and original use of the documents. Provenance and original order also allow for "inheritance" of higher-level metadata meant to describe the home agency from which the documents came and the way the documents were originally arranged.

Finally, an archival approach to curating a collection of Web documents—focusing first on aggregates (collections and series), rather than

on individual documents—trims the number of items that need to be appraised by a human down to a more manageable number.

Arizona Model Summary

The Arizona Model uses a methodology that applies both human selection and automated capture to the archiving of Web content. In this approach, Web materials are managed in a way similar to the organization of materials in paper-based archives: as a hierarchy of aggregates rather than as individual items. This approach reduces the problem of the sheer volume of preserving Web materials to a more manageable size, while maintaining a scalable degree of human involvement. It is the guiding model for OCLC's Web Archives Workbench.

A Tour of the Web Archives Workbench Tool Suite

The Web Archives Workbench Workflow

The Arizona Model is particularly instructive in its evocation of where, in the practice of archival management, automation can be considered most useful. That is, while technology may be applied for information processing activities such as data searching and tracking, and list construction and classification, tasks for distinguishing whether content is in-scope or is valuable are best reserved for humans. Thus, a key deliverable of the ECHO DEPository project, with OCLC as the technical lead, was to develop a suite of tools that would follow the Arizona Model and thus achieve a productive complement between automated processing and human decision making, all the while adhering to established archival principles.

Prior to tool design and development, OCLC carefully considered the user community's needs, which OCLC identified as a blend of librarians and archivists. Significant to its consideration was the issue of terminology: how should tools and features in the Web Archiving Workbench be named if a mixed community of librarians and archivists was to serve as its user base? The word *series*, for example, while familiar to an archivist, might invoke semantics and usage that is different, even unfamiliar, for a librarian. Thus, in exploring the user community, OCLC had archivists look at new types of metadata and asked librarians to think about principles of archiving. Eventually, OCLC elected not to devise new terminology for the concepts at issue; not only did the team conclude that terminology was, in essence, a training matter, it also saw that the work of librarians and archivists often overlap—that is, each is frequently engaged in the milieu of the other.

The software that OCLC created, the Web Archives Workbench (WAW), comprises five tools to identify, select, describe, and harvest Web-based materials, as well as to keep track of, or log, these activities and to generate reports about them. In doing so, they serve as a conduit between human involvement (via manual selection) and computerized capture of Web content: they convert the archivist's policies for collecting content

created on the Web to software-centered rules and configurations. They also assist information professionals by providing the means to add metadata to harvested objects as aggregates. In addition, the tools implement the PREMIS-based METS profiles developed by ECHO DEP at the University of Illinois for packaging content; by design these profiles facilitate ingestion into multiple external repositories and support long-term preservation.[1] Packaging is the last step in the WAW workflow, after which the objects are ready for ingest into an external digital repository (Figure 1).

Furthermore, in doing high-level analysis for the user interface, OCLC arrived at several working assumptions that influenced the design of the tool suite. One assumption was that because the tools in the Web Archives Workbench might change over time, they needed to be "aware" of each other and enable the sharing of data, but—as important—the user should have the ability to opt *not* to use a tool in the Workbench. Through interviews with librarians and archivists, OCLC also learned that harvesting responsibilities often were shared among individuals; as a consequence, data generated by a tool had to be rendered shareable by multiple users—and simultaneously so. This feature would allow a user to view the work of another. In addition, rather than trying to integrate the Workbench into an institution's many authentication schemes, OCLC incorporated a simple scheme, allowing the Workbench to run with just basic administration. In terms of harvesting, OCLC designed more than one harvesting workflow, so that a user could select the appropriate level of analysis and sophistication for a task. For instance, the Quick Harvest feature is a single-screen

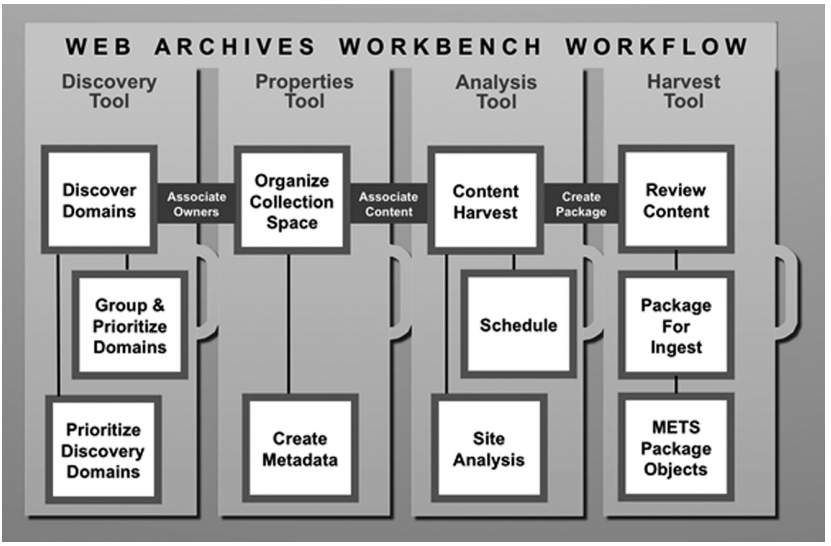

Figure 1. Diagram of the Workflow Encompassed in the Web Archives Workbench

launch point that runs a harvest immediately. The Analysis tool, which is part of an extended harvesting workflow, requires more set-up, but it results in a bigger "pay-off" in terms of the website change observations it handles automatically for the user.

Finally, where the deposit of harvested information is concerned, OCLC recognized that ingest to a variety of repositories, including its own Digital Archive as well as DSpace repositories, would need to be accommodated. A clean, simple interface was created between the point where the Workbench ends and a repository software application would begin; that is, the Workbench generates harvested packages of content in a file system that the repository then picks up and processes. This is the point in the workflow at which the above-mentioned PREMIS-based METS profiles developed by ECHO DEP is implemented.

A Tour of the Software

The screenshot in figure 2 below displays the main WAW tools screen after the user has logged on. The five tools in the Workbench are the Discovery, Properties, Analysis, Harvest, and System tools. In the screenshot they are exemplified by the topmost row of tabs. Although the Alerts tab sits in this row, it is less a tool than a feature of the Workbench. It enables users to access a collection of reports and alerts for the Discovery, Properties, Analysis, and Harvest Tools. In the interface for the WAW tools, a tab is colored in to signify which tool is open, or active, at that particular moment. In figure 2, for example, the Discovery tab is shaded, because the Discovery tool is currently active. Similarly, the Entry Points tab is shaded, because it is active as a component of the Discovery tool.

A key advantage to the Workbench tools is that harvesting of Web content may be scheduled so that it occurs on a regular basis. However, the Workbench tools also offer users the alternative of running a one-time harvest. This is known as the Quick Harvest, accessible via the Harvest tab. Quick Harvest is addressed briefly in the discussion below of the Harvest tool.

The Discovery Tool: Finding Web Content of Interest

The first step in constructing an archive of Web-based resources is to determine which parts of the Web hold desirable, and thus collection-worthy, content. This step lies at the crux of the Discovery Tool. The Discovery Tool aids in identifying potentially relevant websites by crawling relevant "seed" entry points to generate a list of domains to which the "seed" sites link. (Note: An entry point is a specific website URL where the Discovery Tool will begin to search for domains or collect Web content. A domain is a server on the Internet that may contain Web content and is identified by a high-level address. For example, http://www.illinois.gov/news/ is a website, and its domain is "Illinois.gov." (Domains do NOT include "http://".)

In an approach that effectively borrows from citation analysis, the Discovery Tool is designed on the idea that on-topic sites likely point to other

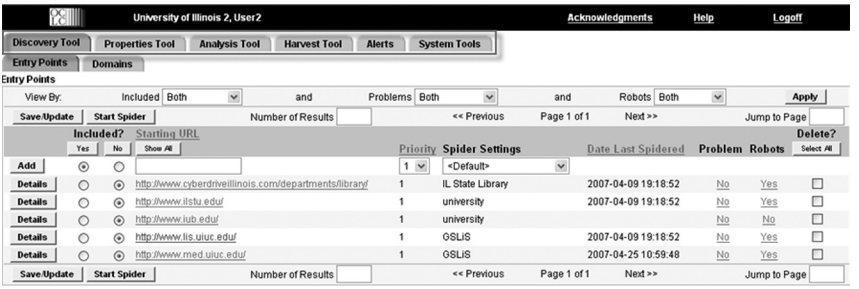

Figure 2. Screenshot of the WAW Interface That User Sees After Logging On

sites addressing a similar topic. The domains in the generated list are then manually evaluated as in-scope or out-of-scope, based on subject interest and collecting policies. Figure 3 shows a list of domains returned after entry points have been crawled, as well as radio buttons that note the scope for each domain. This process results in a list of domains defining a subset of the Web that is relevant for the user's archiving purposes. Domains marked as in-scope can be associated with an Entity (i.e., creator, or agency, or organization responsible for the Web content). Later, in the Properties and Analysis Tools, metadata associated with entities (creators such as agencies or organizations) can be inherited by content harvested from a particular website.

In short, the Discovery Tool is used to:

- generate a list of potentially relevant domains by crawling seed sites;
- assign domains as in-scope or out-of-scope;
- add domains manually to the Domains list;
- associate domains with entities (creating agencies or organizations).

The Properties Tool: Entering Metadata to Describe Content Creators (Entities)
Another premise of the Arizona Model is that, as much as possible, metadata should be entered only once and be inherited by associated harvested objects. After the Entry Points and Domain features of the Discovery Tool are run, and entities (i.e., content creators) have been associated with domains, metadata about the resulting entities may be entered via the Properties Tool. Besides enabling the management of information about entities, the Properties Tool also allows the user to describe the relationships (e.g., parent/child) of entities with one another, as well as enter other information such as contact information.

Importantly, the Properties Tool also can be easily engaged to create analyses and series from entities' websites. The purpose of enabling analysis of a website is to examine its structure—that is, the directories that make up the website. (For more on the Analysis Tool, see below.)

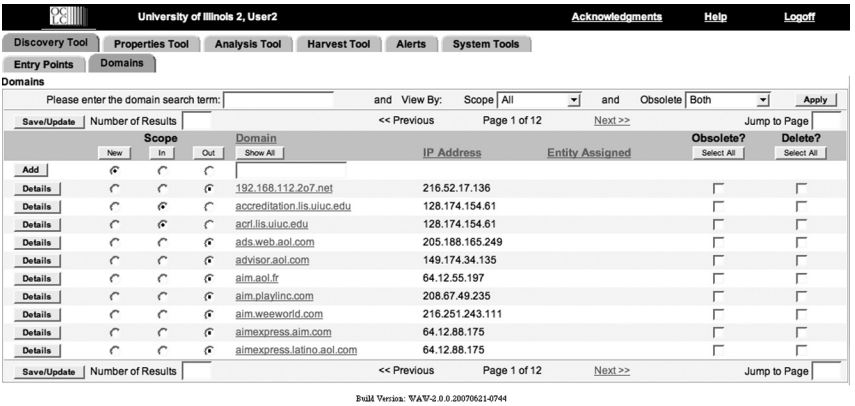

Figure 3. Screenshot of the Interface for the Domains Feature of the Discovery Tool

The Properties Tool is used to:

- create and manage a list of content creators (entities);
- assign metadata and other properties to entities;
- specify websites that entities are responsible for, and create analyses and series based on those websites.

The Analysis Tool: Visualizing the Structure of a Website

Through the Analysis Tool it is possible to discern whether there is valuable content in the directories that comprise a website and, if so, to identify those chunks of content. "Series" refers to flexible aggregates of content that are analogous to archival series—which may be a whole website or a portion of it (e.g., only PDFs of annual reports), or even one individual page or document from websites. Loosely defined, a series is any collection of Web material that a user chooses to collect in one "bucket." In addition, series are used in order to drive the Workbench harvest operations. While series may be established within the Properties Tool, they can also be established and managed using the Analysis Tool, then harvested and packaged in the Harvest Tool.

The Analysis Tool has two functional areas:

- The Analysis screen, which provides visualization tools to aid in content selection decision-making and in series structure decisions. Here, too, a baseline analysis can be created against which to measure future website analyses.
- The Series screen, where series are created, edited, and managed; Series objects are kept; and Series harvests are regulated.

The Analysis Tool is used to:

- analyze the structure of a website;
- enter associated entities;
- set a baseline analysis for comparison with future analyses;
- adjust settings, such as spider settings and change notification threshold settings;
- define a "series" for harvesting (e.g., harvest as an individual object), with option to associate it with an entity;
- hold series objects prior to harvest;
- schedule harvests of series.

In addition, operations for holding series objects and harvesting them may be accessed via the Properties Tool.

The Harvest Tool: Reviewing, Packaging, and Ingesting Harvested Content

All the harvests in the Workbench, including series harvests (via the Analysis Tool) and quick harvests, are listed in the Harvest Tool. The Harvest Tool is used to monitor the status of harvests and to provide an opportunity to review and modify the harvest before packaging it up and ingesting it into a repository. There may be single-object harvests or multiple-object harvests, depending on whether the option to harvest content as individual objects was selected in the Series details screen of an Analysis-based Series (i.e., in the Analysis Tool). The Quick Harvest feature schedules one-time harvests of content based on a URL inputted directly into the Harvest Tool.

After harvests are complete they may be reviewed, at which time additional metadata may be assigned. The user can render, or display, the harvested content within the WAW tool from the Harvest Results page. The user can actually "step into" the harvested content at both the harvest starting point and at any other point in the website (via the website file structure display), and the software will render the website appropriately. The purpose of the display feature in the Web Archives Workbench is to allow the user to verify the correctness of what was harvested—"correctness" meaning that all the information expected to have been collected *is* collected. Once the harvested content is confirmed as correct, it then can be ingested into the user's local repository. Display of content for end users should occur in the local repository. OCLC did not want to duplicate the functions of a repository as part of this project, and it realized that institutions would already have a significant investment in the repository of their choice. This way, users can leverage their existing repository software, with its existing indexing, collection organization, metadata and display functions, and operational (back-up) procedures. The actual code for a repository to display a Web document was not included in the scope of this project.

In sum, the Harvest Tool is used to:

- monitor the status of harvests scheduled in the analysis tool;
- delete completed harvests;
- review completed harvest content, whether single-object or multi-object, prior to ingest;
- review completed harvests; if desired, edit metadata and/or include/exclude content;
- ingest harvested content into a repository;
- launch a one-time quick harvest using the Quick Harvest Tool.

The Alerts Tab: Workbench Notifications

As mentioned above, the Alerts tab is not a tool but, rather, a feature for notifying the user of a variety of systems information. This information includes notification about errors, incomplete processes, completed processes, and new information such as the discovery of a new domain, or a new folder encountered during analysis. In short, the Alerts Tab is used to review reports and alerts about Workbench functions.

The System Tools: Monitoring and Managing Workbench Activities

The System Tools tab contains a number of behind-the-scenes functions that affect and report on activities of the five main tools of the Workbench.

The System Tools are divided into four functional areas:

- The Audit Log page, which displays recent Workbench activities and events;
- The Spider Settings page, where the user can configure default Domain, Analysis, and Harvest spider settings, as well as create additional Domain, Analysis, and Harvest spiders with custom settings. Specifically, types of spider settings include, but are not limited to, depth (how deeply a website should be crawled, or spidered) and parameters of time (when, how frequently, and for how long);
- The Import/Export page, through which the user can import or export a variety of metadata commonly used in the Workbench. These include entities, domains, and subject headings;
- The Reports page, which generates printable reports on activities of the main five Workbench tools. It offers a view of in-development entity and series reports.

Web Archives Workbench Tools Summary

The Web Archives Workbench implements an archival approach to the selection and preservation of digital Web-based content. The Workbench automates much of the methodology embraced by the Arizona Model,

particularly beyond the initial selection decisions made by the archivist (e.g., deciding at the start of the archiving process which website, or which part of a website, to capture and preserve). After selection parameters are set, the Workbench facilitates the capture and management of the digital materials in hierarchical aggregates, not unlike the archiving of print-based materials (See Table 1).

Table 1. Tools Summary

WAW Tool	Purpose/Functionality of Tool
Discovery Tool	Comprising the Entry Points and Domains tabs, the Discovery Tool helps to identify potentially relevant websites by crawling relevant "seed" Entry Points to generate a list of domains that they link to. At the end of this process, the users have a list of domains that defines the subset of the Web relevant for their archiving purposes. From here, the Properties and Analysis Tools are used to manage creator information about domains, and associate this information with harvests of content.
Properties Tool	Comprising the Entities tab, the Properties Tool is used to maintain information about content creators or "Entities" (e.g., government agencies), and associate them with the domains and websites they are responsible for. The Properties Tool also allows users to describe the relationships (e.g., parent/child) of Entities with one another, as well enter high-level metadata about them that may be inherited by content harvested from their websites. Importantly, the Properties tool can also be used to create and associate Series with Entities' websites. Series and harvests are then further managed using the Analysis and Harvest/Package Tool.
Analysis Tool	Comprising the Analysis tab and the Series tab, the Analysis Tool provides website structure visualization tools to aid content selection decisions, and allows users to define archival Series, associate metadata with these series, and schedule recurring harvests of Web content. Harvesting activities are then monitored and managed in the Harvest Tool.
Harvest Tool	Comprising the Harvester and Quick Harvest tabs, the Harvest Tool lists all harvests within the Workbench, including Series harvests scheduled using the Analysis Tool as well as Quick Harvests. It is used to monitor their status, initiate the final harvesting and ingest steps for the completed harvests tracked in the Harvest Tool, including reviewing harvest contents and metadata before ingest. This is the final step in the Web Archives Workbench workflow. It also offers a separate Quick Harvest feature.
System Tools	The System Tools manage and monitor Workbench activities, reporting on operations undertaken in the four other tools. It has four functional sections: an Audit Log page (shows recent Workbench activities); a Spider Settings page (parameters for spidering may be set here); an Import/Export page (for moving metadata); and a Reports page (for producing printable reports about activities performed by the other tools).

Behind the Scenes: OCLC's Technical Implementation of the Web Archives Workbench

An ISO 9001 company, OCLC has an externally audited quality system based on the requirements of ISO 9001 as an aid for ensuring that products meet user expectations and specified requirements. OCLC's project development lifecycle is a process that specifies how OCLC services are marketed and developed. This process includes lifecycle documents such as project plans, requirements, design, test plans, operations support plans, and post-project reviews. The Web Archives Workbench program followed this lifecycle.

OCLC software development teams are free to follow different methodologies within the framework of the OCLC lifecycle. The WAW development team used Dynamic Systems Development Methodology (DSDM), a comprehensive framework for agile project delivery (http://www.dsdm.org). The DSDM methodology applied to many parts of the project, including the requirements-gathering approach, requirements prioritization, and task scheduling. The core development team consisted of a total of four to six developers, two product managers, and one test analyst. Supporting this team within OCLC were systems and network engineers, quality assurance staff, operations staff, and other groups. UIUC provided project management, requirements input, documentation, engineering support, and test beds for the METS-based inter-repository data exchanges.

The WAW program was divided into three main projects and many smaller releases in order to reduce risk and to create a feedback loop allowing refinement of the requirements based on previous releases. There were three major software releases, plus approximately twenty additional releases over the course of the three-year program. The three main development projects were based on the main areas of functionality of the tool suite: (1) Domain and Entity, (2) Analysis and Packager, and (3) Site Analysis and Change Management. Though the Domain and Entity features in WAW were somewhat functionally simple, the Domain and Entity project carried a significant amount of risk because it built the technical foundation on which the rest of the project would rest. The Site Analysis and Change Management tools were risky due to the usability issues involved in clearly representing to the user the process of harvesting and evaluating changes to websites. Throughout the project one of our main concerns was how to represent the Arizona Model in a clear and usable way in software.

Based on early discussions, the system began to be seen as a "workbench," into which components and systems would be incorporated and dropped over time—perhaps because users would prefer to apply some of their local tools or perhaps because they would have multiple tools for a given task. Additionally, each component would grow its data quality over time, thereby forcing the rest of the system to adapt easily to evolving specifications and data versions. Therefore, the architecture is designed

for location, interface, and data-exchange transparencies, which means that changes in those three main areas are expected to drive all other system characteristics.

The high level technical architecture of the system was specified using the Reference Model—Open Distributed Processing (RM-ODP). This framework uses various views of a system, including a domain model view, an information view, an application view, and a technology and deployment view.[2] Using this framework, OCLC created the following early domain model of the system. (See fig. 4.) Some of the boxes in this domain model were later removed from the requirements, as our understanding of the system to be built changed over time.

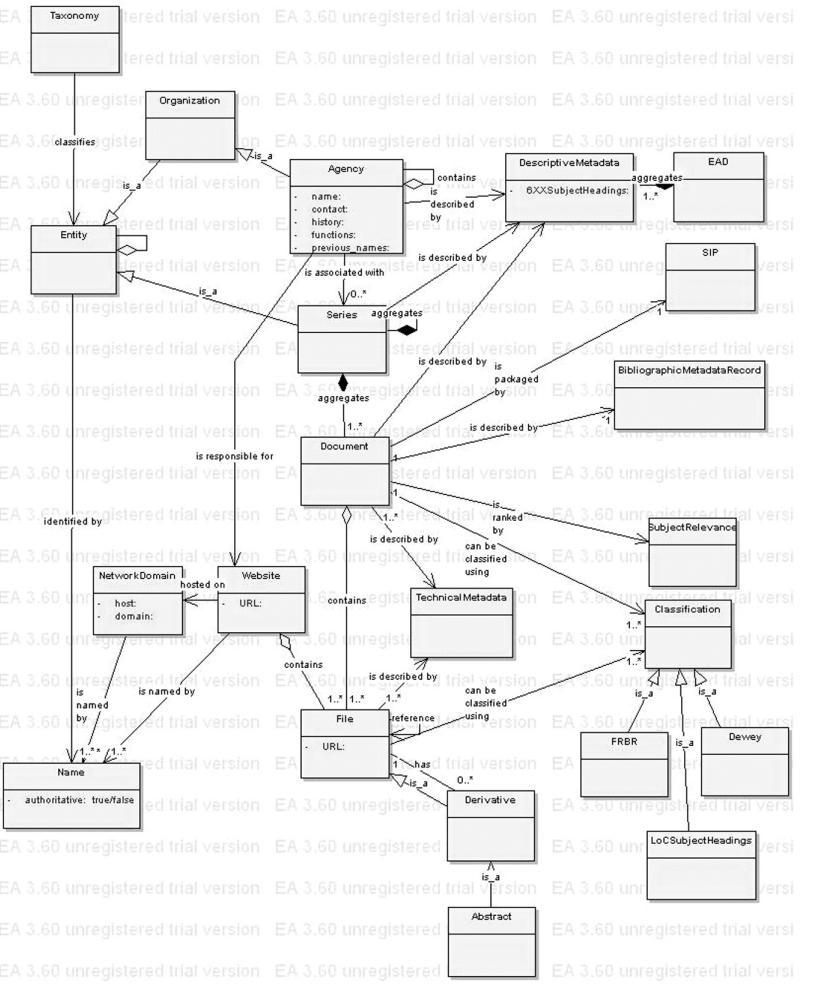

Figure 4. Diagram showing OCLC's early domain model of the system that eventually developed into the WAW suite of tools

The architecture consisted of several layers: client, integration, service, and persistence. The client layer consisted of a user interface implemented using the Struts framework as a model-view-controller structure to the code. The second layer is a Web services layer that provides the hooks for a client to talk to the application. This layer also provides integration between tools and translation between the internal and external representations of the data. Each developing WAW tool (Entity, Analysis, Domain, etc.) implemented a consistent Helper API to allow the user interface layer to Add/Update/Delete/Search single or multiple objects. The Oracle database provided a persistence layer. Once the high level design was produced, a detailed design was produced for each tool. OCLC created use cases for all main activities in each of the tools. The OCLC lifecycle requires formalized review and sign-off of requirements and design documents. Following DSDM meant that detailed requirements and design were produced as needed before each implementation time box started, as opposed to the traditional *waterfall* software development approach where all requirements are written, then all designs produced, then all coding completed.

Each developer worked in his own *sandbox*, where a WAW interface was set up for his exclusive use. The work of multiple developers was integrated into a development test environment called *Baseline*. This way, product managers and test analysts could review work in progress in Baseline. When Baseline was ready it was migrated into a quality assurance environment, where formalized testing was done against a test plan. For major installs, Baseline was also installed at UIUC for additional testing. The final step of the development process was to deploy the software into a production environment.

The Web Archives Workbench was released as an open-source package on SourceForge in October 2007. Release documentation includes detailed installation instructions and a detailed user guide for understanding and using the tools.

- WAW Release home page: https://sourceforge.net/projects/webarchivwkbnch/
- Administration Guide: https://sourceforge.net/project/showfiles.php?group_id=205495
- User Guide: http://is.gd/gKlz
- WAW software package: http://webarchivwkbnch.cvs.sourceforge.net/webarchivwkbnch/webarchivwkbnch/

The Administration Guide has runtime environment requirements for WAW. It also has a list of all third-party software used by WAW in the incorporated code section of the document. The third-party software is included in the WAW distribution. An OCLC subscription is not required to use WAW or to use this third-party software.

The WAW tools, as developed by this project, will continue to be made publicly available indefinitely through SourceForge. In addition, in 2008 OCLC released a new array of services incorporating components of the WAW tools into a workflow with CONTENTdm, WorldCAT, and the OCLC Digital Archive.

Findings—User Feedback

Testing of the WAW tools was undertaken in varying degrees by the original project content partners, as well as by several volunteer organizations. Feedback about their experiences working with the tools was gathered during large-group project meetings at OCLC, as well as through phone conversations and e-mail exchanges. The overall response indicates that the Web archiving approach of the WAW tools was "elegant" and worth consideration, but in practice content partners generally did not implement the full functionality of the tools. Thus, the potential benefits of applying an archival approach to the Web were not realized completely. Reasons for this partial implementation have to do with inadequate resources and time needed for training in the proper use of the tools, which also points up their complexity. The Web Archiving Workbench is powerful and extensive in terms of Web harvesting and content, or series, analysis, but—according to the feedback from our content partners—at a cost of heuristics and usability. Not surprising, the Quick Harvest functionality was engaged most often; for some, the Quick Harvest feature became a much-valued component of their daily workflows. Changes in content delivery approaches—such as from static Web pages to database-driven pages—constituted another reason for not applying the full functionality of the tools.

Limited Resources and Limited Time

During their participation in the ECHO DEPository project, state library and archives partners remained under continual operational pressures to respond to the need for capturing content from agency websites. Some partners tested the WAW tools while continuing to use other Web content capture approaches in order to meet their immediate obligations, leaving fewer resources to focus on the WAW tools. Because the tools were still under development, testing of the various phased releases may also have been difficult to incorporate into daily workflows. Support from the project, in the form of interns, had been planned but was geared to the early releases of the Workbench, before the full functionality of the tools was implemented. In hindsight, putting project resources toward direct work with content partners, as originally intended, might have resulted in more use of the full functionality of the tools, especially if timed more specifically to coincide with later, more fully functional, software releases.

Complexity of the Tools

According to user feedback, the Quick Harvest and Discovery tools were easiest to use, because they could be set up quickly and incorporated into existing workflows without increasing the need for new resources. The full functionality of the tools involves understanding a process with a greater level of complexity than that presented by the Quick Harvest option. Partners reported that it was easier to use the Quick Harvest and Discovery tools rather than expend time and resources for learning, or testing, the tools suite as a whole. Further, some content partners report that the complicated interface of the tools was a barrier to using them to their fullest potential.

Web Content Delivery

The assumption proposed by the archival model—that a website and its directories are similar to an archival record collection and set of record series—does not apply today as easily as it did when the model was first proposed in 2003. An increasing amount of content is now delivered through database-driven websites rather than through static Web pages. The relationships between content items that may have been obvious when stored in a file directory are not always apparent when stored in a database. Therefore, crawling domains to find potential content to harvest and applying inherited metadata according to a directory structure are now less useful approaches than they were just a few years ago. Despite this shift in how information is delivered via websites, the concept of content inheriting metadata from previously harvested content, and then associating that content with an existing aggregate collection, continues to be useful for making automated harvest processes more effective.

Conclusions and Next Steps

State librarians and archivists continue to search for the best methods for capturing Web content based on their specific mandates and the resources they have available to them. Recent developments in Web archiving services and tools provide new opportunities for partnering with others and for exploring new workflows. The Web Archives Workbench tools are one option among many. They automate the methodology prescribed by the Arizona Model, which is premised on key archival practices, such as observation of provenance and adherence to original order. The four main tools (Discovery, Properties, Analysis, and Harvest) enable the identification, selection, description, and packaging of digital content. In addition, the WAW suite includes functionalities for error notification, as well as system tools for overseeing and reviewing Workbench activities. The lessons learned from developing the Workbench, and the underlying archival model used to direct its development, underscore the merging roles

and responsibilities of archivists and librarians in the digital environment and the need to re-evaluate and re-envision workflows.

Moreover, the continuing mission and significance of this work have been affirmed in the second phase of NDIIPP. The University of Illinois, OCLC, and the University of Maryland have partnered to develop a standalone, open-source metadata extraction tool intended to provide access to archived content—a kind of next step for the Web Archives Workbench. In addition, in the State Initiatives component of NDIIPP, a selection of state libraries across the nation are collaborating to develop tools and service models for the management and preservation of state government digital materials. These projects address digital preservation in a variety of contexts, including disaster readiness and the recovery of data. Through the State Initiatives work, NDIIPP is addressing the fundamental issue of keeping at-risk state government resources viable as part of our national heritage and record.

Notes

1. Two METS profiles developed by ECHO DEP are at work here: the ECHO DEP Generic METS Profile for Preservation and Digital Repository Interoperability (2005) and the ECHO DEP METS Profile for Web Site Captures (2006). The former is the "top level" format-generic profile, which focuses on implementing PREMIS. The latter, a Web capture profile, is an example of a "sub-profile," which is used with the first one to provide a structure for more format-specific information.
2. In RM-ODP the architecture of a system is described by five views (essentially five different points of view) reflecting the separation of responsibilities between business sponsors, developers, and support staff. Those views are:
   - Enterprise—community, enterprise objects (domain model), objectives (requirements/use cases), roles
   - Information—schemas, object attributes, data boundaries, constraints, semantics
   - Computational—components, interfaces, interactions, contracts
   - Engineering—transparencies (location, access, failure, persistence), nodes, channels
   - Technology—technologies and products (the only dependence on specific products and implementation packages)

References

Cobb, J., Pearce-Moses, R. & Surface, T. (2005). ECHO DEPository Project. In *Archiving 2005: final program and proceedings, April 26, 2005, Washington, D.C.* (pp. 175–178). Springfield, VA: The Society for Imaging Science and Technology, 2005. Retrieved July 5, 2008, from http://www.ndiipp.uiuc.edu/pdfs/IST2005paper_final.pdf/

ECHO Dep Generic METS Profile for Preservation and Digital Repository Interoperability. (2005). Retrieved August 27, 2008, from http://www.loc.gov/standards/mets/profiles/00000015.html

ECHO Dep METS Profile for Web Site Captures. (2006). Retrieved August 27, 2008, from http://www.loc.gov/standards/mets/profiles/00000016.html

The ECHO DEPository: An NDIIPP-Partner Project of the University of Illinois at Urbana-Champaign with OCLC and the Library of Congress. (n.d.). Retrieved July 5, 2008, from http://www.ndiipp.uiuc.edu/

The ISO Reference Model for open distributed processing—An introduction. (1996). Retrieved August 27, 2008, from http://www.enterprise-architecture.info/Images/Documents/RM-ODP2.pdf

OCLC Digital Management Services. (2008). Retrieved July 5, 2008, from http://www.oclc.org/us/en/services/collection/default.htm

The National Digital Information Infrastructure and Preservation Program. (n.d.). Retrieved July 5, 2008, from http://www.digitalpreservation.gov/

Pearce-Moses, R., & Kaczmarek, J. (2005). An Arizona Model for preservation and access of Web documents. *DttP: Documents to the People.* 33(1), 17–24. Retrieved July 5, 2008, from http://www.ndiipp.uiuc.edu/pdfs/azmodel.pdf/

Rani, S., Goodkind, J., Cobb, J., Habing, T., Eke, J., Urban, R. & Pearce-Moses, R. (2006). Technical architecture overview: tools for acquisition, packaging, and ingest of Web objects into multiple repositories (poster). *Opening information horizons: 6th ACM/IEEE-CS Joint Conference on Digital Libraries: June 11–15, 2006, Chapel Hill, NC, USA: JCDL 2006/sponsored by ACM SIG on Information Retrieval, ACM SIG on Hypertext, Hypermedia and the Web, IEEE Technical Committee for Digital Libraries* (pp. 360–360). New York: ACM, 2006.

Web Archives Workbench. (2008). Retrieved July 5, 2008 from http://sourceforge.net/projects/webarchivwkbnch/

Web archiving. (2008, August 21). In *Wikipedia, the free encyclopedia.* Retrieved August 27, 2008, from http://en.wikipedia.org/wiki/Web_archiving

---

Patricia Hswe is project manager for NDIIPP Partner Projects at the University of Illinois at Urbana-Champaign. In 2004–6 she held a CLIR postdoctoral fellowship in scholarly information resources at the university's Slavic and East European Library. While a student at the Graduate School of Library and Information Science in 2006–8, she worked in the Mathematics Library and in Grainger Engineering Library Information Center; in addition, she was a graduate research assistant on NDIIPP during its first phase. Her research interests focus primarily on digital libraries and digital collections: metadata (standards, creation, semantics, usability, management); the challenges of digital preservation; use and users of digital resources; data curation in the humanities; and information literacy and services for graduate students.

Joanne Kaczmarek is associate professor of library administration and archivist for electronic records at the University of Illinois at Urbana-Champaign. Kaczmarek was a co-PI on the university's NDIIPP project and is involved in several ongoing practical initiatives related to information management and digital preservation. Her research interests include exploring the interplay between technology and human behavior as it relates to the management of information resources and what becomes the lasting, historic record of an individual, a group, an organization, or a culture. Prior to her current position Joanne was the project coordinator for the Mellon-funded OAI-PMH Cultural Heritage Repository project at the University of Illinois.

Leah Houser is a manager in product development at the Online Computer Library Center (OCLC). Houser was the project manager for the OCLC staff who implemented the Web Archives Workbench. Activities for the project included prototyping, requirements, design, programming, testing, product support activities, and communications. She has worked on OCLC's digital preservation products since 2000. Prior to her current position she worked on OCLC's FirstSearch product as an administrative and project manager and software developer.

Janet Eke served as project coordinator of the NDIIPP-funded digital preservation projects based at the University of Illinois at Urbana-Champaign from fall 2004 through August 2008. She is now the research services coordinator at the Graduate School of Library and Information Science (GSLIS) at the University of Illinois at Urbana-Champaign where she helps to develop services, tools, and resources to support and promote the research efforts of the school. Previously at GSLIS she provided research services at a fee-based custom research unit and taught a master's course in online searching. Before joining GSLIS in 1998, she worked for many years in public libraries.