AN ARIZONA MODEL FOR PRESERVATION AND ACCESS OF WEB DOCUMENTS

Richard Pearce-Moses, Director of Digital Government Information, Arizona State Library, Archives and Public Records, rpm@lib.az.us

Joanne Kaczmarek, Archivist for Electronic Records, University of Illinois at Urbana-Champaign Archives, jkaczmar@ux1.cso.uiuc.edu

Originally published in *DttP: Documents to the People* 33:1 (Spring 2005), p. 17–24.

ABSTRACT

This paper describes a model for curating[1] a collection of Web documents. The model is based on the observation that the organization of Web sites parallels the organization of an archival collection and on the assumption that archival principles of provenance and original order are useful to curate and to provide access to documents in the collection. The University of Illinois and OCLC has received a National Digital Information Infrastructure and Preservation Project from the Library of Congress to test and refine the model and to develop open-source software tools to support its practical application.

BACKGROUND

Like many other state libraries, the Arizona State Library is the official depository for state agencies' official reports and publications ('documents'). The Legislature mandates the Library

to collect and provide access to these documents for current and future use. In an effort to reduce printing costs, the Arizona Legislature encourages agencies to publish documents on the web, rather than in paper. As a result, the number of documents published on the Web has skyrocketed. On average, state agencies' Web sites contain more than 300,000 documents at any given time.

In many ways, the Web can be a boon for the State Library. The increased number of documents on the Web means a vastly richer collection of publicly available reports and publications, and the Web has made it much easier to locate and capture fugitive documents. However, Web documents present a number of challenges to traditional ways of curating a print-based collection. The Web is used to distribute ephemeral documents, in addition to official reports and publications, blurring the distinction between which documents should be added to the depository program and those with limited value. Web documents often lack the formal elements of printed reports and publications; without a cover sheet or title page, finding the information necessary to describe the documents can be a challenge. Where printed documents have a simple and familiar structure – ink on paper sheets with a binding that defines the content's sequence and boundaries – Web documents are often created using specialized software and may contain links that blur the document's boundaries.

To realize the potential benefits of the web, the State Library must discover ways to identify, select, acquire, describe, and provide access to the enormous amount of digital information state agencies are now producing. *What* we do will remain fundamentally the same, but *how* we do

those things in a digital environment will change significantly. Those changes are reflected in the vernacular as the words *publication* and *document* are replaced by *information*.

To date, institutions building a collection of Web publications have generally followed one of two models. The "bibliocentric" model is based on a 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. This approach can capture a low volume of high quality content. However, it cannot be scaled to the massive numbers of Web publications without a large increase in human resources.

The "technocentric" model focuses on software applications that can capture virtually everything with automatic Web crawls. This approach 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 content. This approach essentially transfers the work of selection from the libraries to the patron.

A third approach relies on Web masters to send documents to the State Library or to create documents using standards and metadata so that the Library can easily harvest them. Unfortunately, the Library has not had success with this approach. Web masters are very busy and cannot easily take on additional work. Second, they are not trained in collection development, so they are often unsure what documents the Library wants. Finally, because turnover among state Web masters is relatively high, institutionalizing these practices is difficult.

Ultimately, it has always been the Library's job to do these tasks, and the Library must discover another approach that works for the web.

An Archival Approach

The Arizona State Library is investigating another approach to curating collections of Web publications. This model is based on the observation that a Web site is similar to an archival collection. Both are collections of documents that have common provenance. Both group related documents together; on the web, the groups are called directories and subdirectories, while in archival collections they are called series and subseries.[2]

At a fundamental level, an archival approach to curating a collection of Web documents shares the same basic functions as a bibliographic approach. Both involve identification and selection (what archivists call appraisal), acquisition, description, reference and access, and preservation. However, the archival approach to accomplishing those functions is distinguished from a bibliographic approach by a few core principles.

- Materials are managed as a hierarchy of aggregates: collections, series, subseries, and folders. In general, archivists do not manage collections at the item level unless the individual items are of great importance.
- Respect for provenance requires that documents from one source are not mixed with documents from another source.
- Respect for original order requires that documents be kept in the order that the creator used to manage the materials.

- Respect for provenance and original order ensures that documents remain in context, and that the context can yield a richer understanding of the individual documents. The whole, by nature of the interrelationships between the individual items, is greater than the sum of its parts.

The benefits of an archival approach to curating a collection of Web documents, focusing first on aggregates (collections and series), rather than on individual documents, reduces the size of the problem to a more practical number. Spending just five minutes each to process the 300,000-plus Arizona Web documents would take twelve years to complete. Taking an archival approach by spending ten hours analyzing the series (directories) on the 200 collections (Web sites), the work could be done in a year. To the extent series are stable on a Web site, the amount of work after the initial analysis will be substantially less in subsequent years.[3]

These archival principles also facilitate efficiently curating the materials in the collections. Exploiting original order saves time spent reclassifying materials in another system. The principle is based on the assumption that the creator needed to be able to find materials and that order remains useful. If the Department of Transportation organized its construction records by highway, reorganizing the records by contractor, date, or some other facet will not necessarily improve access.

These principles also facilitate efficient access to the materials in the collections. Working with aggregates does not necessarily reduce efficiency of patron access. Large aggregates (collections) are organized in a manner that makes it easy to eliminate large quantities of

irrelevant materials. Someone looking for information on highway construction would start with the agency principally responsible for that activity, the Department of Transportation. Provenance can help identify other collections that might have relevant information; Environmental Quality may have some information based on the impact highways have on air and water pollution. Many times provenance can help eliminate large bodies of material from a search; Child Welfare Services probably has little or no information on highway construction.

Smaller aggregates (series) are also organized to help narrow a search even further. The patron looking for highway construction within the Department of Transportation might disregard series associated with driver's licenses and traffic studies. The series on highway construction presents the patron with a rich collection of relevant materials that, when supported with appropriate finding aids, makes it easy to find the precise information being sought.

The Craft of Curating a Collection

Curating a collection of Web documents using archival principles is relatively straightforward. The archivist approaches the documents on a Web site as an organic whole, then, moving down the hierarchy, looks at each series in the collection as a whole. The archivist stops when further subdivision of the hierarchy is no longer useful.

The challenge of curating a collection of Web documents is in understanding the structure of the Web site. In particular, the archivist may have access to the documents through the Web site, but may not have direct access to the underlying server or its file system.

Specialized software can facilitate the process of curating a collection of Web documents as an archival collection. However, the tools alone will not guarantee success. First and foremost, the Arizona Model focuses on craft rather than technology. It seeks to articulate a rational way to perform tasks and to use tools in an integrated fashion to produce a reasonable result. Having a hammer and chisel does not make one a sculptor, and few can use those tools to create a work of art.

IDENTIFICATION

The first step in building the collection is determining which Web sites have documents appropriate for acquisition. For the Arizona State Library, those boundaries can be easily described as "All Web sites that hold state publications." At first glance, identifying the specific URLs for those Web sites might seem as easy as including all domains that end in state.az.us. In fact, while conventions used to organize the Web offer clues to the intellectual organization of the content, the two are far from congruent. One of the principal tasks of identification is to describe the boundaries of state agency Web sites in terms software can understand.

'Web site' is an ambiguous term. In many cases, a Web site correlates with a single domain; for example, www.azexample.gov hosts all content produced by the (fictitious) Arizona Department of Examples, and a link to each page on that Web site begins with www.azexample.gov. In the case of large Web sites, the boundaries may encompass several domains, which may be clearly associated, such as photos.lib.az.us and www.lib.az.us. Often, a large Web site may integrate content from very different domains, such as www.azdeq.gov and www.phoenixvis.net.

The boundaries of Web sites are often blurred by the relationship between different content providers[4] that are responsible for different portions of a Web site. The Division of Paradigms, a part of the Department of Examples, publishes a set of documents and services on the Web that would be commonly considered the Division's Web site. However, the relationship between the Division's and the Department's Web sites may not be as clear and their organizational relationship. The Division's Web site may be nothing more than a directory under the Departments (www.azexamples.gov/paradigms) or it may be part of a sub-domain (paradigms.azexamples.gov), or it may be a complete different domain (www.azparadigms.gov).

The concept of Web site is further confused by the complex relationships between content and creators, and by the lack of consistent practices for organizing material on the web. As a result, only a human can define a Web site's boundaries. The Arizona Model tests machine-assisted approaches that combine the strength of computing with human judgment, to identify and maintain a list of domains (and corresponding Web sites) with in-scope.

Machine-Assisted Identification of Web sites

The first approach is based on the assumption that the vast majority of state Web sites will be referenced on at least one other state Web site. By analyzing the links on all state pages, it should be possible to discover the links for those Web sites. Starting with a seed list of Web sites, a spider downloads all the pages on those sites and builds a list of all links on those pages. The list is then analyzed to create a list of distinct domains. For example, a Web site may include many links to different pages on the state Department of Transportation Web site, each of which begin with the domain www.dot.state.az.us. The site may also include many links to

pages on the federal Department of Transportation Web site, which begin with www.dot.gov. In Arizona, the initial scan of four large Web sites captured some 10,000 links, but less than 700 domains.

The list of domains was then manually appraised, and if the domain represents or is part of a state Web site, it is associated with the content provider. Many domains were immediately recognizable as out-of-scope (www.dot.gov, www.adobe.com) or in-scope (www.dot.az.state.us, www.azgovernor.gov). Sites that staff was not familiar with had to be evaluated manually, a process that took about three days.

As each new state Web site is discovered, it is added to the seed list so that it will be included in future analyses. However, the software will only show newly discovered domains, so after the initial evaluation the list of domains to be reviewed will be relatively small.

Manual Identification of Web sites

The machine-assisted approach is an efficient, but imperfect tool. While there is a high correlation between Web sites and a single domain, that correlation is imperfect. This approach cannot distinguish Web sites located on commercial servers (www.mindspring.com/~abote). Also, this approach does a poor job identifying subordinate Web sites for divisions within an agency (the Arizona Capitol Museum, a part of the State Library, has a Web site – www.lib.az.us/museum – that is part of its parent's Web site).

To counter these limitations, documents with lists of agencies, such as organizational charts, budgets, and telephone directories, are manually searched to identify names of entities not already discovered using the machine-assisted tool. Those names are then searched using Internet search engines to discover if the agency has a Web site that has not yet been identified. This second approach helps catch exceptions the first approach misses, including Web sites that are not referenced on any other state site and Web sites hosted on commercial Internet service providers.

Reconciling Web sites with Content Providers

Building the list of domains is merely a means to an end. The ultimate goal is a list of content providers and their Web sites. Each domain is associated with a content provider, and the content providers are organized into a taxonomy that documents the relationships between content providers (agencies with their subordinate divisions) and that links content providers to their Web sites.

Table 1. Taxonomy

Library and Archives	www.lib.az.us
Archives	www.lib.az.us/archives/
Photo Collection	photos.lib.az.us/
Braille and Talking Book	www.lib.az.us/BTBL/
Law and Research	www.lib.az.us/is/
Genealogy	www.lib.az.us/is/genealogy/
Museum	www.lib.az.us/museum/
Transportation	www.azdot.gov

Aviation	www.azdot.gov/aviation
Motor Vehicle Division	www.azdot.gov/mvd
	www.servicearizona.ihost.com
Planning Division	tpd.azdot.gov

The taxonomy will also include descriptions of the content provider (described below in Description and Access) so that patrons can quickly see if a Web site is likely to have relevant information.

The taxonomy will never include every agency, division, department, office, board, commission, and task force. The Arizona Model relies on human judgment to determine when it is appropriate to create an entry. While Genealogy is listed in the example above, in practice it may be sufficiently described in the entry for the Law and Research Library. Building the taxonomy will take time. However, it can be grown incrementally, starting with little more than the content providers' names and later adding in additional information.

The taxonomy database is one of the key tools of the Arizona Model, recording information about and providing a systematic view of both the content providers and their Web sites.

SELECTION

Once a state Web site has been identified, the second step is to determine which documents on that Web site should be acquired for the depository program. Using an archival approach, selection is done at the series level, rather than considering each document individually.

An archival series – not to be confused with the bibliographic concepts of series and serials – is "a group of similar records that are arranged according to a filing system and that are related as the result of being created, received, or used in the same activity."[5] The Arizona Model's presumption that series exist on Web sites is founded on the common human behavior of organizing related materials into groups to help manage them. Because this is a general behavior rather than a requirement, different Web masters will organize their sites differently and with varying degrees of consistency. Those idiosyncrasies mean that a series-level approach to selection will have varying degrees of success. At one extreme, sites may lack any order. However, larger sites tend to have at least some order.

For example, a Web site may have a series (directory) called 'forms;' looking at a sample of documents, it quickly becomes clear that the series contains blank forms that are outside the scope of collections. Another series called 'calendars' contains documents about upcoming events and, because of their ephemeral nature, are outside the scope of collections. A series called 'reports' contains documents that are clearly within the scope of collections. Because series names may be often misleading, a human must sample documents in the series to see if the documents in the series are in scope. A series called 'annual reports' looks like a rich find for the depository, but on closer inspection the documents are blank forms used to file annual reports with the agency.

Of course, many series contain a mixture of documents, only some of which are in scope. In many cases, a decision would typically be made to collect all or none on the grounds that there is not sufficient staff time to select individual documents.[6] As a result, some out-of-scope

documents will be preserved or some in-scope documents will be lost. In some instances, it may be possible to write rules to exclude unwanted documents or to include in-scope documents. For example, a directory containing both pdf and html versions of the same document might have a rule to capture only one format.

Series are often broken into subseries. For example, a series called reports might be broken into subseries for different years or for different geographic areas. That practice is also reflected in some Web masters' organization of their sites, with directories broken into subdirectories to further organize content. In the case of very large Web sites, the hierarchical structure is very deep, extending to subseries within subseries within subseries. The art of archival selection is knowing when to stop traversing the hierarchy. To use the example above, the subseries for years or regions need not be evaluated separately if the larger series is selected.

In some instances on the web, a directory is not a series, but represents a single document. The subdirectory – sometimes called a folder at this level – contains different files that form the document. This use of directories is especially common at the lower levels of the file structure. The Arizona Model assumes that most selection decisions will be made at a higher level.

In order to be able to appraise and select at the series level, archivists need site analysis software to help them visualize and understand the directory structure of a Web site, especially if they do not have direct access to the system. For example, an analysis of the links on one Web site revealed more than 7800 links to files on the site, although those files are organized into a total

208 directories, subdirectories, and folders, but only the top 31 directories are equivalent to series appropriate for selection.

When archivists understand a Web site's structure, it is possible to make decisions about selection, including series to avoid and how often to acquire documents within specific series. The site analysis software will record these selection decisions.

Description and Access

The Arizona Model envisions a combination of traditional archival description, high-level manual indexing using a controlled vocabulary, and full-text indexing to facilitate patron access to the documents in the collection. The model envisions search software that groups documents into categories for more precise results and to facilitate discovery by helping patrons narrow their search.

In archives, description and access are hierarchical. Patrons use collection-level descriptions of content providers (provenance) to determine which collections are likely to hold documents relevant to their interests. For example, someone searching air pollution would start with the Department of Environmental Quality before the Department of Economic Security. Once collections are identified, patrons look at descriptions of the series to determine which likely hold relevant materials. In the case of air pollution, the patron would disregard series on water quality and landfills. The patron can also take advantage of any subseries; in this example, the patron might look at a series on auto emissions. After locating relevant series, the patron browses a list of the documents in those series.

Traditionally, archivists have described each collection in a finding aid. A finding aid usually begins with an administrative history, which explains the content provider's mandate and functions, and curatorial information, such as acquisition information and restrictions. This information is captured in the taxonomy database. The heart of a finding aid includes a scope and contents note, which describes the nature and type of the collection as a whole; an outline of the series and subseries, which serves as a rough table of contents; and then a list of the folders, organized by series and subseries. This information is captured in the site analysis tool. Patrons browse the finding aid, much like looking over the captions on the folders in the files, and request only those files they are interested in. The Arizona Model adapts this process with two modifications.

First, the collections and series are assigned descriptive metadata from a controlled vocabulary;[7] when documents are harvested from the series, that metadata is assigned to each document within the collection or series. The Department of Water Resources might be given the headings "Water Resources." The Governor's Drought Task Force, a division within that department, would be given a narrower headings "Drought" and "Water conservation." The series "County programs" on the task force's Web site might be given the heading "Planning documents."

Second, the finding aid does not list folders within series, but the titles of documents. In the context of manual preparation of finding aids, listing each document in an archival collection was prohibitively time consuming. By harnessing the power of the computer, it is possible to create a list of all documents. The quality of that list will vary with Web masters' individual

conventions for naming documents, as well as the ability of the computer to distinguish documents.

While a finding aid's bird's-eye view of a collection remains a useful curatorial tool for archivists, patrons would almost certainly find finding aids cumbersome access tools in the age of full-text searching. The Arizona Model sees the descriptive metadata as a tool to help display the results of full-text searches more efficiently by organizing the results into categories. A full-text search would retrieve all documents that contained the search words, and then sort the results under headings drawn from the controlled vocabulary in the descriptive metadata. The search would then allow patrons to narrow their search to just documents in those categories.

For example, a categorized search for water might look something like the following, which would precede the more familiar item-level results organized by a relevancy-ranking algorithm.

Table 2. Hypothetical Results from a Categorized Search

Found documents in the following categories		
water (500+)	water conservation (357)	Salt River Project (210)
drought (110)	flood control (98)	xeriscape (25)
Found documents in the following categories		
Water Resources (135)	Drought Task Force (102)	Phoenix (87)
Maricopa County (84)	Corporation Commission (35)	Rural Watershed Alliance (76)

Even though the public is familiar with full-text search engines like Google, they are often frustrated and stymied by results that bury useful documents under thousands of false hits. Categorization has proven value in making searches more efficient. Based on an analysis of global information locator service (GILS) searches at Washington State and here in Arizona, people greatly prefer searching by browsing categories. Browsing categories is particularly

valuable because it helps ensure that searches are constructed using the terms used in the documents. Browsing also prevents patrons from having to think of the right concept; rather, they are offered choices to help them find just what they're looking for.

The taxonomy database records information about the content providers, and the site analysis software records metadata about the series.

ACQUISITION

In traditional archival and bibliographic workflows, acquisition takes place before description. In the Arizona Model, acquisition is an automated process that follows rules developed during the preceding steps. Using information about which series contain documents to be acquired for the depository, harvesting software downloads documents from those series. It creates a package that contains all the files necessary to reconstruct the document. Descriptive metadata taken from the document's parent collection and series is added to the package. Administrative and preservation metadata is also added to the package. Additional descriptive metadata may be added using text analysis software. That package can then be stored online or offline. The harvesting software may also insert text into the package so that patrons viewing the document clearly recognize that it is an archival document with historical, non-current content.

Harvesting software will likely use the Metadata Encoding and Transmission Standard (METS) as the basis of the package structure. Separate software will be developed to load the package into digital collections software, such as Greenstone, Fedora, DSpace, and OCLC's Digital Archive.

THE REALIZATION OF THE ARIZONA MODEL IN THE OCLC UIUC NDIIPP GRANT PROJECT

In September, the Library of Congress announced that it had awarded the Library and the Graduate School of Library and Information Science at the University of Illinois at Urbana Champaign a grant under the National Digital Information Infrastructure and Preservation Program (NDIIPP). Other partners in the grant include OCLC, the Arizona State Library and Archives, the Connecticut State Library, the Illinois State Library, the North Carolina State Library, the Wisconsin State Library, the Tufts University Perseus Project, and the Michigan State University Library. A significant part of the project will be to test and refine the Arizona Model and to develop open source software tools to support the practical implementation of the model.

---

[1] Throughout the paper, 'curate' is used to refer to a wide range of activities, including identification and selection, acquisition, description, and reference and access.

[2] Materials of the same provenance are sometimes called a collection, a fonds, or a record groups. The words are not exact synonyms and each reflect variations in archival theory and culture. For the purposes of this paper, these approaches are more common than different, and for the sake of simplicity the authors use the word 'collection' throughout.

The divisions within archival collections may be given different names in different contexts. Large collections in large repositories often have more divisions, including subgroup, series, subseries, and subsubseries. Large Web sites may have similar divisions, with agencies using different domains, directories, subdirectories, and subsubdirectories. Throughout this paper, 'series' and 'directory' will be used to describe the internal divisions

within a collection or Web site; 'subseries' and 'subdirectory' will be used only when essential to emphasize the notion of a series within a series or a directory within a directory.

[3] The time to process the documents would be 1,500,000 minutes or 25,000 hours at 60 minutes/hour. Given 2,080 hours in a work year, it would take twelve years of uninterrupted work to finish, without consideration for any new documents that might be added. Evaluating large collections (Web sites) – a process discussed later in the paper – will certainly take more than the suggested average of ten hours. However, many smaller collections have simple structures with few directories and will take far less time. With some small Web sites, it will be more efficient to capture everything than to spend time analyzing them.

[4] 'Content provider' is an unfortunate bit of jargon. Necessity demands its use as a generic term to represent the diverse names given different levels of government, including agency, division, department, office, board, commission, court, individual, or other body responsible for providing Web content.

[5] Pearce-Moses, Richard. *A Glossary of Archival and Records Terminology* (Society of American Archivists, forthcoming 2005). Pre-publication copy available from www.archivists.org/glossary/ [accessed 15 December 2004].

[6] Preliminary specifications for the site analysis software allow selection decisions to be made at the document level, should a repository have the resources to work at this level. Even repositories with limited resources may want to take advantage of this feature for some series, where the time spent weeding will have sufficient benefits.

[7] Many state libraries are using a controlled vocabulary developed by Jessica Milstead for this purpose. Available from www.cyberdriveillinois.com/library/isl/lat/lat_findit_subject_tree.html [accessed 15 December 2004].