



Collection Plan
for
Arizona State Agencies Web Publications

10 January 2007

Prepared by:

Richard Pearce-Moses
Arizona State Library, Archives and Public Records
rpm@lib.az.us

Contents

Section 1.	Mission & Scope	3
Section 2.	Selection	5
Section 3.	Acquisition	7
Section 4.	Descriptive Metadata	8
Section 5.	Presentation and Access	9
Section 6.	Maintenance and Weeding	11
Section 7.	Preservation	11
Appendix A.	Submission Agreements	12

Section 1. Mission & Scope

A. Mission Statement

The Arizona State Library and Archives is a library of many special collections, including the state's law library; unique, historical maps; genealogical materials; published Arizoniana and Arizona imprints; archival records, personal papers, and historical photographs; and a professional collection. The Library and Archives is a regional federal depository library.

Territorial and State Agency Publications is one of the most significant collections, which is the core of the State Agency Publications Depository Program. The State Agencies Web Collection represents the transformation of those materials into digital formats. The State Agencies Web Collection supports each of the State Library's key goals.

Preserving Arizona

From the state's beginnings, Arizona has made provision for the State Library to build a collection of agency publications.¹ The founders of the territory recognized the importance of these reports as what we would now call corporate memory, allowing those currently serving in government to understand the historical context of their circumstances: a record of the state's decisions and actions and the underlying rationale. The importance of capturing state agency publications was reaffirmed in the Constitution when Arizona became a state, and again in 2006 when the Legislature passed legislation giving the State Library additional authority and power to acquire state agency publications.

Providing Access

The value of any collection is seriously diminished if the materials are inaccessible; the real value of these materials is in their use. The ability to make electronic publications widely accessible to Arizonans – and the world – through the Internet greatly enhances their value. Further, the Internet provides the State Library an opportunity to fulfill its mandate in a way that was impossible in a time when people had to travel great distances to access government information housed in a central location.

Prompt, professional service to the Legislature

The State Legislature is a primary constituent for this collection. The Legislature (including both the elected members and staff) benefits through direct access, as well as through the State Library's reference and research services.

Promoting collaboration

The State Library and Archives recognizes that it cannot achieve its goals in isolation. Building this collection allows it to develop relationships with the agencies publishing materials on the web. The State Library and Archives coordinates its efforts to curate this collection with other libraries that collect state government information.

¹ Howell Code (23 HC 1864), Sec. 14. There shall be and hereby is established a Territorial library, to be located at the capital of the Territory, and the members of both houses of the Legislature, and the executive and judicial officers of the Territory shall, at all times, have free access thereto, under such regulations as shall be made by the secretary of the Territory, who is hereby the Territorial librarian.

B. User Groups

The audience likely to use the collection is similar to the core audience of the State Library: the Arizona Legislature and Arizonans. The primary value of the collection reflects the purpose for which the records were made and provides evidential value regarding Arizona government. The secondary value of the collection reflects information for other uses. Arizona history is, without a doubt, a principal subject area, as are law and genealogy. Given the breadth of government the collection will provide significant information that provide historical context for nearly every discipline, ranging from anthropology to zoology.

C. Collection Subject, Theme, or Event

As mandated by law (ARS 41-1335(B)) the State Library shall:

1. Acquire and provide access to materials relating to the following in print, in an electronic format or in any other format: a) law; b) political science; c) economics; d) sociology; e) subjects pertaining to the theory and practice of government; f) genealogy; g) Arizona history.
2. Provide the following: a) a general and legal reference service; b) a records management and archives program; c) a state and federal government documents depository program; d) a library development service; e) museums for educational purposes as approved by the board; f) a service, including materials, for persons who are visually or physically unable to use traditional print materials.

Further, statute (ARS 41-1338(A)) states, "The state library shall contain . . . 2. Copies of current official reports, public documents and publications of state, county and municipal officers, departments, boards, commissions, agencies and institutions, and public archives."

As used in the collection development plan, "agency" is defined as every state office, whether legislative, executive, or judicial, and all of its respective officers, departments, divisions, bureaus, boards, commissions, and committees, all state-supported colleges and universities which are defined as state institutions of higher education, and other units, such as societies, which expend state-appropriated funds.

"State publication" incorporates works in any format that may or may not be financed by state funds, but which are released by private entities pursuant to a contract with or subject to the supervision of any agency, such as research and consulting firms under contract with a state agency. "State publication" does not include ephemeral materials, correspondence, intra-office or inter-office memoranda, routine forms, or "records" as defined in ARS 41-1350. Superseded material (publications containing information cumulated in later issues, issued in later revised editions, or replaced by other volumes) are excluded from the collection.

D. Curator(s)

The Director of the Law and Research Library is primarily responsible for the collection, with input from the Director of the History and Archives division. The Director of Digital Government Information is responsible for managing the technical aspects, including harvesting, storage, and preservation. Day-to-day responsibilities for the collection are assigned to the Digital Collections Librarian.

Section 2. Selection

A. Seed List

Websites targeted as sources for documents are identified using a process described in "An Arizona Model for Preservation and Access of Web Documents."² Briefly, URLs on twelve key agency websites are harvested.³ The URLs are parsed and a distinct list of domains is returned for review. The list of domains to review is many times smaller than the complete list of URLs; 50,000 URLs produce a list of approximately 2,500 URLs. (Ultimately, the project hopes to review all domains on all agency websites.)

Staff can quickly recognize and flag many of those domains as out-of-scope; for example, virtually every website includes a link to Adobe.com, directing people to a source of the Acrobat Reader. Agency staff typically recognize a third of domains as out-of-scope. Similarly, agency staff are familiar with many state agency sites (typically following a format az[abbreviation].gov or state.az.us), and can quickly flag those as in-scope. Roughly a third of domains are immediately recognized as in-scope. The remaining third of domains much be manually reviewed, a process which generally takes only a few minutes.

The most current list of domains for Arizona state agency websites can be found at <http://findit.lib.az.us/agencies.asp>

B. Capture Scope

The collection does not acquire websites as an organic whole. Rather, it collects documents as distinct items. (The acquisition process uses rich metadata to document the context of the document on the website, allowing researchers to know the document's original context, and the researcher can search for other documents within that context.)

The web has blurred the distinction between publications and records. Traditionally, 'publication' referred to a work that was produced in multiple copies, providing broad distribution of unvarying content. A 'record' was typically unique and preserved as evidence of a transaction or to preserve memory of something important.

That distinction has always been fuzzy. Important records were published; the National Historical Publications Commission (now the National Historical Publications and Records Commission) was founded to publish documentary editions of records to provide broader access to those records and, primarily, to ensure their preservation through geographic distribution of copies. Moreover, many publication-like documents were unique or of limited number; for example, dissertations were published by depositing a copy in a library.

The web makes the same content available to all individuals who access it. Hence, by definition, all items on the web are publications.⁴ However, the State Agency Web

² Richard Pearce-Moses and Joanne Kaczmarek. Originally published in *DttP: Documents to the People* 33:1 (Spring 2005), 17–24. Online at <http://rpm.lib.az.us/azmodel/AzModel.pdf>.

³ Arizona State Portal (az.gov/); Arizona Dept. of Administration (azdirect.state.az.us/); Office of the Governor (azgovernor.gov/); Arizona Telecommunications and Information Council (www.arizonatele.com/atic/default.htm); Auditor General (www.auditorgen.state.az.us/); Dept. of Agriculture (www.azag.gov/); Dept. of Environmental Quality (www.azdeq.gov/); Dept. of Transportation (www.azdot.gov/); Secretary of State (www.azsos.gov/); Dept. of Water Resources (www.azwater.gov/); Citizens' Clean Election Commission (www.ccec.state.az.us/ccecweb/ccecays/home.asp).

⁴ As another example of the web blurring these distinctions, what might normally be clearly considered a publication may not be distributed in exactly the same format. Different browsers may

Publications collection does not include content that would traditionally be considered records. Many times, this content is accessed through a web interface to a database, and current spiders cannot download such content.

Using a process of macro appraisal,⁵ the Library and Archives has categorized all state agencies and offices into three levels of collection emphasis: High, moderate, and low. Priority is given to those agencies and offices in this the high category, those that play an essential role in state government and that develop law and public policies affecting the whole of the state. Obvious examples include the Legislature and the Governor; other examples include the Department of Environmental Quality, the Department of Water Resources, and the Department of Transportation. As much attention as possible is given those agencies and offices in the moderate category, which includes [examples]. In reality, little time may be left to agencies and offices in the low category. As these agencies and offices often produce limited content; in these instances, the time spent selecting materials has little benefit in terms of space saved, so all content may be captured by default.

The State Agencies Web Collection excludes some content, such as ephemeral information, blank forms, or calendars of events. However, such information may be acquired from those agencies and offices with the high macro appraisal scores.

If the Library and Archives routinely and reliably received print versions of material published on the web, it may choose *not* to acquire the web version if the content is of limited value or presents particular preservation problems. The decision to not acquire an electronic version is not made lightly, as the ability to provide rapid, broad access over the web is very important. Factors that influence this decision include the ability of the harvesting software to capture the content, the need for access to the web version (as opposed to other versions that may be generally available), and the value of the content. In some instances, the Library and Archives will work directly with an agency to acquire content, especially database-driven content, rather than through the web.

C. Rights Management

Unlike the federal government, the State of Arizona does not have legislation that places state agency works in the public domain. As such, copyright inheres in the publication, and the state is the owner of copyright. Because the State Library and Archives is part of the state government, it has rights to capture and redistribute state agency publications. At the same time, the State Library and Archives recognizes that some state agencies derive revenue from their publications, and the State Library and Archives acquires, but does not reproduce, those publications in a competitive manner.

State agency publications are also covered by public records laws, which grant the public certain rights of access to these works. Embedded materials that are copyrighted by some entity other than Arizona fall under this statute.

State law directs the Library and Archives to establish rules and regulations regarding access to records. Those policies include the requirement that confidential information be protected. In 2006, the state strengthened the protection of confidential information by passing a law protecting confidential information. Courts recognize that some confidential information may appear in unlikely places and that it would be impossible

render the content differently. Moreover, the content may differ from user to user (the server may alter content based on information about the user) or instance to instance (rotating images).

⁵ For more information, see "An Arizona Model."

to review every single publication for such information; access to this information is covered under the principle of practical obscurity and is not considered a violation.⁶

The Library and Archives presumes that agencies have already screened information on their website so that no confidential information is disclosed. At the same time, if staff discovers a series of publications that contain confidential information, access will be restricted as defined by law.

Section 3. Acquisition

A. Frequency of Capture

The Library and Archives seeks to spider agency websites on a monthly basis. The agency will encourage agencies to leave all content on their websites for at least a month so that the Library has a chance to capture that content. If the agency observes that a site changes frequently, especially an important site, it may choose to spider the site more often. In the case of exceptional events that trigger rapidly changing content on agency websites, it may spider all or some sites on a frequent basis.

B. Capture Scope

The list of websites to spider is repeated monthly. Because the staff maintains a database of domains they have reviewed, they need only review newly discovered domains. The monthly review process also flags domains that are no longer discovered, allowing staff to identify obsolete domains.

As noted above, the State Library does not seek to capture all content on agencies' websites. Rather, it seeks to select documents that would normally be acquired in paper format. Given the enormous number of documents on agencies' websites, item-level selection is not possible.

Following the Arizona Model, staff will take advantage of the structure of agencies' websites to select documents in aggregates defined by the structure of the website. The Web Archives Workbench abstracts a website's directory structure by analyzing the links on that site. This approach is based on the archival principle that record creators organize similar materials to facilitate managing those records. For example, a directory on the Dept. of Water Resources groups information relating to the Governor's Drought Task Force (in scope), while another directory on the Corporation Commission's site contains nothing but blank forms (out of scope).

Although a directory may include some out-of-scope materials, such as ephemeral information, time spent weeding those materials is not justified by the costs of storing and providing those materials. Contrawise, a directory that contains largely out-of-scope materials may include a few items that should be acquired, and the Library recognizes that some in-scope materials may be lost.

C. Material Types & Formats

The Library seeks to capture state agency publications in all formats. Publications received in non-standard formats will be copied into a standard format supported by the Library and stored with the original. Currently, Library regulations specify Adobe Acrobat

⁶ See "practical obscurity" in Richard Pearce-Moses, *A Glossary of Archival and Records Terminology* (Society of American Archivists, 2005). Online at http://www.archivists.org/glossary/term_details.asp?DefinitionKey=3053.

(PDF) as the standard format. The Library is investigating standards for formats that PDF cannot adequately support.

D. Interactive & Dynamic Content

Ideally, the Web Archives Workbench will flag forms-based pages for careful review. Because spiders typically cannot retrieve information from form-based sites, the librarian must assess the best approach for that content.

A significant amount of the “deep web” on agencies’ websites falls under the statutory definition of record, and is outside the scope of the collection. However, if the records merit long-term preservation, they would be transferred to the State Archives through the existing records management program.

If the information constitutes a publication and the spider can crawl the site (all content is accessible through links), the spider is given hints as to how to capture the site. If the spider does not work correctly and the content is sufficiently valuable, it may be captured manually. If the publication (or publications) cannot be easily crawled, the digital librarian will contact the agency to negotiate transfer of the publications using some other method.

Section 4. Descriptive Metadata

A. Level of description

All agencies and their subordinate units are described at a level appropriate to their importance and size. A major agency may have divisions and offices larger than many other state agencies. Agency descriptions include

- Website creator’s name, including official (legal) form, standard forms (AACR2), and a “key name” that places the central concept in the name in the first (for example, the Dept. of Transportation’s keyname is Transportation Dept.). The cataloger may assign as many variants as is appropriate to aid retrieval.
- LC Name Authority ID.
- Local identifying numbers. (AzDOC classification number)
- Legal mandate establishing the agency.
- Dates the agency was established or disbanded.
- Predecessor or successors.
- A narrative administrative history⁷ that explains these data elements in terms understandable to humans. “Typically such histories are not exhaustive but usually include only information that is relevant to the materials in the collection. They commonly include the full, legal name of the organization, as well as common short forms and earlier and later forms of the name; prominent dates, such as of charter, mergers, or acquisitions; function and mission; position within the hierarchy of a complex organization, including parents and subsidiaries; principal officers; and places of operation.”
- Associated names: Typically used to record incumbents of key offices, such as governor or secretary of state. May be used to record the head of the agency.

Each website will be described as a collection, generally in accordance with *Describing Archives: A Content Standard*. Series (web directories) to be harvested will be

⁷ From Pearce-Moses, *Glossary* (administrative history). Online at http://www.archivists.org/glossary/term_details.asp?DefinitionKey=505

described hierarchically; if a series description is sufficient for subseries, the subseries will not be described separately.

B. Metadata elements

Currently, the Library and Archives staff are reviewing the Global Justice XML Data Model and the National Information Exchange Model for “universal” metadata elements common to the vast majority of government records. Currently identified minimum metadata elements include:

- **Creator.** Collections: based on the agency responsible for the site (or portion thereof); includes authority form of the name and may include the AzDOCS classification number. Documents: Inherit creator metadata from collection and series plus any internal metadata.
- **Title.** Collections: typically a standard for “Agency name: \$bweb documents, \$finclusive capture date.” Document: Based on internal metadata and machine analysis of document.
- **Date of capture.** Collections: described in terms of the range of dates, typically when capture was begun and frequency of capture. Documents: described in terms of date captured and dates the same document was found on the web.
- **Extent.** Collections and series: Probably none. Documents: file type.
- **Administrative history.** Collections and series: narrative written by the cataloger. Documents: None.
- **Scope note.** Collections and series: Narrative written by the cataloger. Document: May be autogenerated, but at least one scope note will include the first 1,024 characters of the document to support full-text indexing.
- **Subject headings.** Collections and series: Assigned by cataloger. Documents: Inherited from collection and series level descriptions, plus autogenerated headings and Dewey Classification.
- **Hash value.** Documents only: MD5 or other hash value to determine if document has changed.
- **Context.** Documents only: The original URL, as well as the collection and series (and subseries) where the document was located.
- **Structural information.** Documents only, as necessary.

C. Controlled vocabularies

Metadata assigned by catalogers at the collection (website) and series (directory) level will conform to LCSH and AACR2.

Section 5. Presentation and Access

A. Discovery

Users will be able to do a full-text search of the first 1,024 characters of a document.⁸ Ideally, this full-text search will return results categorized by agency and

⁸ Research indicates that a full text search of the first 1,024 characters is only slightly less precise than a search of the entire document.

subject metadata.⁹ Users will be able to browse the collection with finding aids generated by the accessions database that records the acquisition of the documents.

B. Access

The vast majority of the collection will be open. As noted above, series likely to contain confidential information will not be made available. Series descriptions will be made available so that the public will know that the records exist and can take actions to gain authorization for access.

Agency publications that are commercialized will not be accessible as long as they remain available for purchase from the agency of origin.

C. Look-and-Feel

As much as possible, the look-and-feel of the original should be maintained. However, it should be clear to all users that they are accessing non-current documents. This notice might be accomplished through a variety of techniques, including click-through pages that users see before they reach the publication or a banner supplied at the top of a window that frames the publication's content.

In some instances, it may not be possible to maintain the look-and-feel of a document. The digital librarian should determine a format most appropriate to preserve the look-and-feel while ensuring, first and foremost, that the content is preserved as nearly perfectly as possible.

D. Dynamic Content

Not all links in harvested documents will work, and some links will work unpredictably. To the extent possible, links within the document should remain active and point to content archived at the same time. Links outside the document should be disabled, although users should be able to easily discover and navigate to the URL the link originally referenced; before leaving the document for the external link, the user should be informed that the content referenced may have changed.

E. Multiple Types/Formats

As much as possible, the public will be allowed access to all content. Although the Library may not be able to render the bitstream properly, the user may have the tools or be willing to build tools to render the bitstream.

F. Authenticity

The Library and Archives may need to certify the publications' authenticity. To do that, the Library and Archives will rely on the trustworthiness of the system, demonstrated by established policies and procedures; an accessions log to demonstrate chain of custody; and documented preservation routines and hash values to demonstrate integrity.

⁹ See "An Arizona Model" for a fuller discussion of the desired retrieval model.

Section 6. Maintenance and Weeding

A. Maintenance Activities

Seed lists will be updated monthly based on new domains discovered using the Web Archives Workbench.

B. Deselection Guidelines

It may not be possible to render many formats in the future, making the materials inaccessible. However, the Library recognizes that an initial failure to capture content while it is accessible guarantees the material will be lost. In the future, the Library may deselect obsolete materials that cannot be rendered.

Note: For the foreseeable future, the Library will keep all previous versions of the document. This practice is based on the assumption that migration techniques are very new. As those techniques mature and the process becomes easier and more robust, it may be better to base future migrations on the earliest version of a document, rather than a recent version that has been migrated several times. At some point in the future, the Library may choose to deselect obsolete versions of a document.

C. Collection Evaluation

Web logs will be analyzed monthly for access statistics and to determine patterns of use.

Section 7. Preservation

A. Technology Obsolescence

On an annual basis, the digital librarian will run a report of all file formats in the database and make plans to migrate obsolete formats before the Library loses the ability to render those formats. The digital librarian will assess the quality of migration, documenting "quirks." If the publication cannot be rendered after migration, the digital librarian will look for a base-level format (plain ASCII or Unicode, a raster graphic) that will allow the content to be captured at some level.

B. Preservation Metadata

Library and Archives staff are currently reviewing the LMER and PREMIS documents to establish standards for preservation metadata.

Appendix A. Agency Notification

The staff will work with the Director of the Library, who will send the following notice to the director and web master of each agency whose content is harvested. The letter should be sent at the beginning of each fiscal year.

The following model language must be reviewed by the Director before being sent.

The Arizona State Library is mandated by law (ARS 41-1335 and ARS 41-1338) to acquire all state agencies' official reports and publications. This program has been in place since Arizona was formed as a territory.

The law requires the Library to capture these documents, regardless of format. With the advent of the Web, the cost of publication dropped significantly and agencies routinely began publishing many documents on the Web. To acquire these publications in as efficient and effective manner as possible, the Arizona State Library has developed software to harvest publications from your website.

This project will support your agency's compliance with ARS 41-1345 and the Library and Archives' rules and regulations in the Arizona Administrative Code (R2-3-502, R2-3-503), which require your agency to provide the State Library copies of all your publications. The project will also allow you to reduce storage space on your system, while retaining access to historical versions of your web content.

We plan to spider your site on a monthly basis to capture publications. We will not preserve your entire site. To the extent possible, the software will exclude non-publications, such as correspondence, intra-office or inter-office memoranda, routine forms, or "records" (as defined in ARS §41-1350). The software is typically unable to capture information accessed through a form, and it often cannot access information on pages linked using Java script.

We hope that you will take the following steps to improve the quality of the documents we capture.

1. Leave documents on your website for a minimum of one month. This will ensure that the software has the chance to find and capture the document.
2. Allow the robot access to areas of your website that contains information provided to the public. Disallow access to areas of your website that contains no information directly accessible to the public, such as directories for scripts and images.
3. Use directories to group similar materials (something you probably do already).
4. Avoid embedding links in Java.
5. Use meaningful metadata in your documents, recording the office or division responsible for the document, the title of the document, and the date it was issued. This information can be included in the properties page of documents created using Office and most other desktop software or as meta tags in an HTML document.

Our digital librarian [name, email, contact] can answer any questions regarding these suggestions.

Documents captured from your website will be made available to the public via the Internet. The archival copy will be clearly marked as non-current and will include the date the document was captured. The archival copy will also include a pointer to the location of the original on your website. The Library will avoid providing access to any

content that generates revenue for your site as long as that information remains on your site. Please let us know about any fee-based content to which we should restrict access.

Please be assured that the harvesting software is specifically designed to place minimal loads on your site. You will be able to identify the spider by its IP address [list]. Second, the software respects a robots.txt file; the robot's name is [name]. Should you experience any technical difficulties as a result of the spidering software, please contact [name, email, phone number; backup name, email, phone number].

If you have any questions or concerns about this program, please contact me directly (602-542-4035; gawells@lib.az.us) or Richard Pearce-Moses, the Director of Digital Government Information (602-542-4035; rpm@lib.az.us).