From Bibliographer to Curator: Archival Strategies for Capturing Web Publications

Presented at the
International Federation of Library Associations'
Section on Government Libraries Pre-Conference
Oslo, Norway
August 2005

GladysAnn Wells
Director and State Librarian
gawells@lib.az.us

with Richard Pearce-Moses
Director of Digital Government Information
rpm@lib.az.us

Arizona State Library, Archives and Public Records
Phoenix, Arizona, United States of America

INTRODUCTION

In English common-law countries, especially the United States, a covenant exists between the governed and those who govern. That covenant affords citizens legal, judicial, and operational protection and predictability. Citizens expect government to document its actions, defend them, and protect their rights and entitlements. Librarians, archivists, and records managers play an important role in this relationship. They must work with document creators to ensure that the products of governmental business processes are identified and preserved so that they can build a collection of the authentic[1] and authoritative[2] documents and records that allow citizens to hold government accountable and to ensure the legal context within which they work, live, and build their communities.

[1] Authentic documents are genuine, not a counterfeit, and free from tampering. Authenticity is typically inferred from internal and external evidence, including its physical characteristics, structure, content, and context. They are perceived as genuine, rather than as counterfeit or specious; bona fide. Definition based on Richard Pearce-Moses, *A Glossary of Archival and Records Terminology* (Society of American Archivists, 2005). Available online at http://www.archivists.org/glossary/.

[2] Authoritative documents are written or published by an official or organization with specific responsibility for the content of the document.

The rapid rise of e-government and the shift from paper to digital documents has an enormous impact on the library, archives, and records professions. Unless the professions can adapt – as rapidly as the technological environment changes – their role in this covenant will be compromised, and citizens may not have the information they need to protect their rights and interests, nor will government be empowered to maintain its role within the covenant – or defend its actions.

In the twenty-first century, born-digital information does not come in the precise packages with which we were familiar. We must adapt our practices, standards, and services within this new environment. But in order to succeed, those changes must be based on a new understanding of how we *curate* a collection of government information, not just serve as its bibliographer. While *what* we do remains the same, *how* we do that is changing. We can no longer be passive recipients of what we have received through often venerable programs, but must actively capture the information we know will be needed in the future.

I use the word *curate* to refer to a wide range of activities that all cultural professions share: identification and selection, acquisition, description, reference, and preservation.

THE CONTEXT OF OUR WORK

In 2002 Nancy Bolt presented my paper "Needed: Librarians, Archivists, and Records Managers for the Organization, Access, and Preservation of Government Information in the Electronic Era" to this conference.[3] Since then, the pace of change has continued at breakneck speed. In that paper, I opined that technology – and especially born-digital government information – had forced a transformation on the three information professions: librarians, archivists, and records managers. As I said in that paper

> I believe to deal effectively in the digital government environment, practitioners must combine skills heretofore practiced separately by archivists, records managers, and librarians. I believe we must fundamentally re-think the previous, now in the digital era

[3] Presented at the IFLA Section on Government Libraries Pre-conference, 15 August 2002, London. See http://www.lib.az.us/about/pdf/iflapaper.pdf.

almost artificial, differentiation between these skills and build towards new hybrid professionals who can borrow from each of our profession to support the operations of e-government.

The Arizona State Library, Archives and Public Records has increasingly done just this – placing us among the leaders in our professions' struggle with born-digital government information.

Politics are often concerned with the immediate. Politicians must worry about re-election or re-appointment. Similarly, technology is judged mainly by speed of transaction. Technology experts are expected to focus primarily on today's production or necessary results. Conversely, the skills of the records, archival, and library communities focus not only on today, not only on a set of interactions between service provider and service recipient, but also on the documentation that will prevent a repeat of mistakes made yesterday and provide for a future that will last longer than tomorrow.[4] We must believe in the value we bring to the table in the digital age. We must embrace the digital age of government content and develop the expertise that ensures the evolution of our disciplines as practitioners in the government world.[5] Our historic and current contributions – authenticity, authority, and respect for past practices balanced with current usability – are critical in the electronic information era.

We possess the ability to work with all parties and to participate in all types of forums; to be knowledge navigators (the finders of electronic content) in all types of information brokering (business transactions relating to access to electronic content);[6] to establish standards that are used across multiple fields of endeavor; and to teach practitioners and recipients (or users) how

[4] Robert Martin, "The Role of Libraries in Lifelong Learning" (presented at the Five States Indian Conference, 9 May 2002, Mesa, Arizona). Dr. Martin is Director of the Institute of Museum and Library Services.

[5] Michael Totherow, personal communication, 10 May 2002. Mr. Totherow was the Director of Computer Services, Office of the Arizona Secretary of State.

[6] Nancy Bolt describes the services that United States librarians offer government workers in her paper "Serving State government – Librarian Skills, User Education, Services and Support" (presented at the International Federation of Library Associations Section on Government Libraries Pre-Conference, 15 August 2002, London). Ms. Bolt is the State Librarian of Colorado.

to access materials and services. I believe unless the three disciplines (library, archives, and records) work together, collaborate with technology professions, and educate political leaders, the information we require to provide our services, do our work, run our governments, or live our lives, will be lost by accident or design.[7]

It is Our Job

A fundamental tenet of the work we have done in Arizona, through our three professions, has led to what is now called the Arizona model. The Arizona Model, which will be discussed later in this paper, recognizes that our success cannot be predicated on someone outside our own organization. We cannot expect web masters (or others) to send us documents, to remember our job when they are caught up doing their own. We cannot expect them to embed metadata to empower us to serve users they may never meet or report to in any real time manner. We cannot hope for them to provide us unfettered access to their systems. If they do any of these things, all for the better; but we will not presume that we will get their cooperation. In a nutshell, we cannot discipline someone who is not doing something if they do not work for us. And, in reality it is not their job; it is our job. But we can empower them through education and patience to be our partners or at least keep them from being even our passive adversaries.

Long-term preservation must be a part of document creation. While technology offers the promise of information available 24/7, that access will become never/forever if we do not take steps to counter rapid hardware and software obsolescence. Records managers and archivists teach us about the lifecycle of information – what must be kept and what should not, need not be kept. Technology can also help us protect and preserve these materials.

As Dr. Robert Martin, head of the Institute for Museum and Library Services, notes, "If we are to get control over digital records, we need to assure that the information systems are developed to address the needs of records systems. The technology provides ample opportunity

[7] Wells, "Needed," p. 2.

to manage digital records easily, if we only build the capacity into the systems from the ground up, rather than chasing it after the fact." Dr. Martin further notes that in the mid-1990s, Paul Evan Peters, founder of the Coalition for Networked Information urged us to stop standing under the torrent of materials pouring out of the pipeline and trying to sort them out. Instead we need to get up at the top of the pipeline and sort them as they go into it.[8]

We must come to an understanding and acceptance that digital production of information has swamped our capacities to keep everything – even if that were desirable. We must work with all levels of government, bring all skills sets to the table to carefully pick our battles around those items and types of material that must be kept – for the protection and longevity of the covenant – for government to continue to function.

Even as technology has created new, acute expectations and challenges, it has also given us new tools . . . and the capacity to work to develop newer ones for our own skill set needs and practices. Technology allows us to do things that were impractical or impossible just a few years ago. We cannot think of technology as magic. It cannot do everything. We must understand how we can combine the power of technology with human judgment. How do we build systems that take advantage of the speed and power of computers to support human decision making? New tools that are truly effective and efficient are essential to help us work in an environment of reduced budgets.

Technology blurs the distinctions between different types of materials. In particular, we must understand the difference between information and records. Records contain information. But more important, records fix information in a form that is authentic and reliable. We must be able to prove that the records are genuine and without fraud and that we can trust them as accurate accounts of the past – of actions and transactions. Authenticity and reliability are essential to appraisal – the process by which we must decide what is essential to keep and for how long – the

[8] Robert Martin, personal email, 28 May 2002.

lifecycle of information defined through the fixity of records within the context of provenance (where an information set came from, when it was originated, and who worked on it).

As Dr. Martin observed, "While all records systems are information systems, all information systems are not records systems. Records systems require that context and structure be managed as well as content. Information systems usually only manage content." [9]

One of our profession's most embarrassing failures has been our reluctance to rethink how we do things. When we have adopted technology, we did not always pursue business process reengineering. We focused on how technology could help us do the same thing faster, not how it could help us do new things better. For example, when we automated the card catalog we merely put it on technologically-provided wheels, and only later began to investigate new ways to search and display the entries.

The information professions must become even more flexible and adaptable. We are comfortable with well-established standards that evolve slowly. The standard for MARC format (Z39.2) was first issued in 1971.[10] Since then it has become a pervasive technology to support storage, transfer, and retrieval of bibliographic information. But this venerable standard reflects data processing practices of its time. It has lost its monopoly. We now consider other data architectures, such as Dublin Core, Encoded Archival Description,[11] VRA Core,[12] and others yet to be designed.

Arizona Efforts to Curate Collections of Digital Government Information

[9] Robert Martin, personal email, 28 May 2002.

[10] *Information Interchange Format: Z39.2-1994* (Bethesda, Md.: NISO Press, 1994). Available online at http://www.niso.org/standards/resources/Z39-2.pdf (checked 21 April 2005).

[11] Encoded Archival Description (EAD) is a standard data architecture developed by the Society of American Archivists that is used to create finding aids that describe collections of archival materials. For more information, see http://www.loc.gov/ead/.

[12] The VRA Core is a standard data architecture developed by the Visual Resources Association that is used "to describe works of visual culture as well as the images that document them." For more information, see http://www.vraweb.org/vracore3.htm

The Arizona State Library and Archives is involved in several projects to transform bibliographic work to curation of the state documents depository program. Our efforts reflect a cross-disciplinary approach that draws on practices from libraries, archives, and records management enabled by technology capable of operating within the political context and needs of the covenant.

In Arizona, the state's library, archives, and records management programs are in the same agency. We have worked hard to build a team approach to addressing the problems of moving from the world of paper to the world of electrons. We have pooled the insights and expertise of the professions to find novel approaches to traditional practices. Because of our collaborative environment, we were able to see how archival principles of aggregate control could be adapted to curating a collection of state documents published on the web.

The web has been a great boon to citizen access to government information. The web has significantly reduced the cost of publishing information. State agencies can now afford to distribute many more documents than was economically possible in print. As a result, the number of state documents has exploded. At any given time state agency websites hold more than 400,000 documents on more than 150 web servers.

Web SafetyNet

Our first project was in collaboration with the Illinois State Library (ISL). Several state libraries agreed to test software to capture state agency websites. The software was developed for ISL by the University of Illinois at Urbana-Champaign under a U.S. Institute of Museum and Libraries Services (IMLS) grant.

This software takes a techno-centric, brute-force approach to the problem of curating a collection of state web publications. It uses little human judgment in selecting materials, and captures as much as is technologically possible. The software stores the documents in a

compressed format that saves significant amount of disk space, but prevents the documents from being easily accessible. The Library has the documents, should someone need them. However, because Google and other search engines cannot find the pages, the public generally is unaware of them and will not ask for them. The task of finding them remains with the traditional professionals' understanding of what they have – its just electrons, not paper.

For more information on this project, see http://www.isrl.uiuc.edu/pep/.

The Arizona Model for Web Preservation and Access

Given that quantity of material on agency websites, traditional, item-level methods for bibliographic control were impractical. Based on our observation that the organization of websites parallels the organization of archival collections, we developed a model for curating a collection of web publications based on the archival principles of provenance and original order.[13]

The Arizona Model addresses traditional library functions of identification and selection, acquisition, description, reference, and access. However, it looks at those functions through the eyes of an archivist. As I said earlier, *what* we do remains the same, but *how* we do those things is changing. The Arizona Model articulates a different way of accomplishing those things.

Archivists focus on aggregates, not items. The two principal groupings are collections (based on provenance[14]) and series[15] (based on the original order used by the creator). An

[13] The discussion of the structure of archival collections and websites that follows is a generalization. Like collections of paper records, the organization of documents on a website varies from site to site. Some collections are well organized, others less so. It will be harder to apply the Arizona Model to poorly organized sites. However, larger sites tend to be better organized because the web master needs to be able to manage many more documents.

[14] Provenance, sometimes called *fonds*, is "The origin or source of something. . . . Provenance is a fundamental principle of archives, referring to the individual, family, or organization that created or received the items in a collection. The principle of provenance or the respect des fonds dictates that records of different origins (provenance) be kept separate to preserve their context." Pearce-Moses, *A Glossary of Archival and Records Terminology.*

[15] A series 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; a file group; a record series." Pearce-Moses, *A Glossary of Archival and Records Terminology.*

archival collection – one based on provenance – includes materials created or received by an individual or an organization.

In general each agency has its own website, and because all the documents share the same provenance, the Arizona Model treats each website as an archival collection. While collections based on provenance may seem like a coarse sieve, it is surprisingly effective. When browsing a list of collections for information on water pollution, a patron will more likely look at Environmental Quality or Water Resources before they look at Child Protective Services.[16] Managing 150 collections (websites) is much more realistic than managing 400,000 documents. Provenance also supports the authenticity and authority of a particular document. We know where it came from and who is responsible for it (or at least gives it imprimatur by putting it on their website).

Each website organizes similar documents into directories, and the Arizona Model treats each directory as a series. Rather than sorting individual documents into standard categories established by the library (classification), archivists respect original order by keeping the documents in the series established by the creator. Original order makes it possible to refine searches within a large collection based on the characteristics the creator used to organize the documents. For example, someone looking on the Department of Water Resources' website for information on drought would check the series for the Governor's Drought Task Force and ignore irrelevant series, such as those for pollution.

Work priority and processes take advantage of these groups. Websites are ranked in terms of their relative importance using the archival theory of macro appraisal. The most energy is placed on the most important sites. Documents are selected at the series level, rather than the item level. For example, reviewing the series (directories) on a website, it becomes apparent that

[16] Many times, the relevance of a collection to some subject is immediately apparent from the agency name. However, archives typically develop an list of collections with a note about the creators' functions and the scope of the materials to help when relevance is less apparent. For example, it may not be apparent that the Corporation Commission regulates the utility industry. In addition to the note, archives may index the list to help patrons easily find the right agency's materials.

a directory containing blank forms can be ignored, while a series (directory) containing reports must be looked at more carefully.

More information on the Arizona Model can be found at http://www.lib.az.us/DigitalGovt/AzModel/AzModel.pdf.

University of Illinois/OCLC ECHO Depository Research Project

The Arizona Model is a major component of a research project developed by the University of Illinois at Urbana-Champaign and OCLC. The ECHO DEPository research project (Exploring Collaborations to Harness Objects in a Digital Environment for Preservation) is being funded by the Library of Congress through a grant under the National Digital Information Infrastructure and Preservation Program (NDIIPP). The project partners include the Library and the Graduate School of Library and Information Science at the University of Illinois at Urbana Champaign; 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.

The ECHO Depository Project is developing software tools to support an archival approach to curating a collection of digital state documents. The tools are intended to harness the power of computers to analyze the enormous amount of information on the web and present it in a way that supports the work of individuals who use human judgment to curate the collection. For example, people describe groups of documents, and search engines do full text searches to find particular documents within the relevant groups.

Cost Recovery Legislation (HB2187, 47 Leg. 1st R, Ch. 151, Laws of 2005) – A Related Result

Currently, agencies are required to send the State Library a copy of each public report and publication. Compliance with that law is limited because the state documents depository program is not high on their list of priorities.

During the last legislative session the Library was given a tool to help increase compliance. The Legislature, understanding that costs saved by reducing printing, had been transferred to the Library. Therefore, this new chapter authorizes the Library to charge state agencies for the costs of acquiring, describing, and preserving documents. In the process of developing this legislation, the Library had to address two important issues.

First, what constitutes public reports and publications? Before the advent of the web, it was fairly clear what those terms meant. The manner of production distinguished them from records and internal documents that are inappropriate for the state documents depository program. Because the web has significantly reduced the costs of publication, agencies are putting more and more information on the web. As a result, the act of public dissemination is no longer an appropriate measure of what we want. Public access and publication have been conjoined. We had to develop criteria that clarified – both for the agencies and our staff –the kinds of information we wanted for the depository program. See Appendix A for the criteria.

Second, the Library had to plan for success. The Library asked for at least one paper and one digital copy of each publication. The paper copy serves primarily as an insurance policy; we know we can manage and preserve paper over time using traditional skill sets. The digital copy has become the master and is invaluable for providing the kind of 24/7, online access that people expect. As a result, we have had to establish a technological framework to manage and store these electronic documents. We acquired CONTENTdm software and a robust file server to help us manage the documents at the bit level. We have also had to develop standards, policies, and procedures for digital publications. Currently, we are requesting that agencies submit the electronic copy in Adobe Acrobat (PDF) format. The policies, currently under development, are intended to ensure the authenticity and reliability of these documents.

CONCLUSION

We cannot merely tweak our current activities. We must continue to identify and select materials, acquire them, describe them, provide access to and reference service supporting the materials, and preserve them. But, we must think about these activities at a more abstract level and, for the long term, in a significantly different way. That is why throughout this paper I have used the word *curate* to emphasize the changing nature of our work and the similarity among related professions.

To build truly valuable collections that support the covenant between the government and the governed, we must shift our focus from the materials to the context in which the materials are created and the ways in which they will be used. For example, effective cataloging can no longer be measured first or primarily by adherence to rules and items cataloged. Rather, we must return to its original purpose – to ensure that our patrons have access to and can find the information they need, when they need it, and in a form they can use. We must not focus on just the collections but must include the creator as well as the users of born digital information in our solutions.

Appendix A: Requirement to Deposit Public Reports and Publications

COLLECTIONS CRITERIA

Since it was founded in 1864, the State Library has collected and made permanently available to the public official reports and publications for the collective memory of the state and its citizens, so that they have information about public policies and programs – past and present – and to promote government accountability.

To support this program, Arizona law requires all officers and agents of state and local government, including agencies, boards, and commissions, ("agencies") to deposit with the State Library copies of all public reports and publications (ARS 35-103, ARS 41-1335 (B), ARS 41-1338 (2), and ARS 41-4153).

These reports and publications include works, whether in print or electronic formats, that are published, disclosed, or distributed to the general public (or a targeted audience within the general public); and also at least one of the following

- that are required by law as a public report; *or*
- that are required by law to be sent to the Governor, President of the Senate, or Speaker of the House; *or*
- that describe an agency's activities, programs, or policies, including annual reports; *or*
- that are the results of a formal study or investigation.

Shortly after distribution or publication, agencies shall send at least one print copy and one electronic copy (in Adobe PDF) of all public reports and publications to the State Library at no charge. Copies of reports and publications shall include a title page for each document that includes the agency name, title, and date and place of printing or publication. As appropriate, the title page shall also indicate authors, individuals, or organizations that assisted in the production of the report, and any citation to the statute or regulation requiring the report.

This requirement does not include non-public documents, including materials of a confidential nature or materials intended for use primarily within the agency, such as correspondence, forms, interoffice memos, or other materials produced for internal administrative or operational purposes. Non-public documents and agency copies of public reports and publications should be managed according to records retention schedules, which may specify some materials to be transferred to the State Archives at some future date.