ECHO DEPository Project

Judith Cobb
Online Computer Library Center, Inc., Dublin, Ohio, USA

Richard Pearce-Moses
Arizona State Library, Archives and Public Records, Phoenix, Arizona, USA

Taylor Surface
Online Computer Library Center, Inc., Dublin, Ohio, USA

Abstract

The Exploring Collaborations to Harness Objects in a Digital Environment for Preservation (ECHO DEPository) project aims to address the issues of how we collect, manage, preserve, and make useful the enormous amount of digital information our culture is now producing. Collecting, selecting and preserving digital information requires approaches and resources that are substantively different from those we have used traditionally.

The project is a partnership among the University of Illinois; the Online Computer Library Center (OCLC); the National Center for Supercomputing Applications (NCSA); Tufts University's Perseus Project; the Michigan State University Library; and an alliance of state libraries from Arizona, Connecticut, Illinois, North Carolina and Wisconsin.

More specifically, the project will develop criteria for selecting digital material for capture and preservation, with OCLC taking the lead to build software to help automate the process. Illinois, OCLC and NCSA will jointly provide storage for the digital content collected in the project in databases called "repositories" and will test real-world problems that are encountered in the process of digital archiving. The University of Illinois will also conduct research into issues surrounding the long-term semantic preservation of digital resources.

Introduction—The Arizona Model for Managing Web Content

In many ways, the web can be a boon for a library or archives responsible for collecting, managing and providing ongoing access to resources. The increased number of documents on the web means a vastly richer collection of reports and publications, and the web has made it much easier to locate and capture documents that may have never been received in print. 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, making it difficult to clearly distinguish documents that should be added to a collection. 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 collecting organization must find ways to identify, select, acquire, describe, and provide access to the enormous amount of digital information that is now online. *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 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 number 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 web content captured in the process. This approach essentially transfers the work of selection from the curating organization to the patron.

An Archival Approach

The ECHO DEPository project is investigating another approach to curating collections of web publications. The model is based on an approach developed by the Arizona State Library and will be implemented by tools developed by the Online Computer Library Center, Inc.

This model is based on the observation that a website 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.

The approach is based on the following archival tenets:

- Materials are managed as a hierarchy of aggregates. 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 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 web documents would take twelve years to complete. Taking an archival approach by spending ten hours analyzing the series (directories) on the 200 collections (websites), the work could be done in a year. To the extent series are stable on a website, the amount of work after the initial analysis will be substantially less in subsequent years.

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 website 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 the hierarchy is no longer useful.

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

Specialized software can facilitate this 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.

The Web Archives Workbench

The software being developed will facilitate this archival approach to collections of web content. The Web Archives Workbench will consist of the following tools:

Discovery

The first step in building a web collection is to identify the web sites that have content your organization wants to collect. The Discovery Tool will provide machine-assisted identification of web domains that need to be analyzed for potential collecting work.

The tool is based on the assumption that the vast majority of websites will be referenced on at least one other related website. Thus, by analyzing the links on all pages, it is possible to discover related domains. Starting with a seed list of URLs a spider builds a list of all links on those pages, and analyzes the links to create a list of distinct domains. In Arizona, the initial scan of four large websites captured some 10,000 links, but fewer than 700 domains.

The list of domains is then manually evaluated. Some domains will hold content that is within the scope of the organization's collecting policy; other domains will be out of scope. The user will mark each domain as "in scope" or "out of scope" and, based on those indications, the Discovery tool will continue to monitor the list, looking for new domains that need to be evaluated and domains that no longer exist.

Properties

Building the list of domains is merely a means to an end. The ultimate goal is a list of content providers and their websites. Each domain is associated with a content provider, and the content providers are organized into a taxonomy that documents the relationships between content providers and links content providers to their websites. Further descriptive information about the content provider can be added and will be inherited by captured web content using the Analysis Tool.

Analysis

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

An archival 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."[1] The Arizona Model's presumption that series exist on websites 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.

In order to be able to appraise and select at the series level, the Analysis tool will provide a site analysis to help users visualize and understand the directory structure of a website. When the user is able to understand a website's structure, it is possible to make decisions about selection. The user will be able to identify existing series, name each series, describe them, and then indicate specific criteria for how often a spider should search for new and changed content within each series. When the spider identifies new and changed content, the user may choose to be notified, or have the tool automatically capture the content into their collection. Each series is associated with its content provider's properties, and those properties are inherited by the series, and the individual documents within the series.

Packager

Once series have been identified and content within each series is captured and a copy of the content is acquired by the system, the packager tool creates an information package that contains all the files necessary to reconstruct any document within the series. Descriptive metadata taken from the document's parent collection and series is created and administrative and preservation metadata is added to the package.

The Packager tool will likely use the Metadata Encoding and Transmission Standard (METS) as the basis of the package structure. The information packages created will be usable by different types of digital collections software, such as Greenstone, Fedora, DSpace, and the OCLC Digital Archive.

Digital Repository Testbed

As part of the ECHO DEPository project, several organizations will be providing content, including state libraries, which will use the Web Archives Workbench to gather content, for an evaluation of several different digital repositories. As digital content collections grow simply storing digital objects in a computer's file system becomes an inadequate solution to their management.

Large and diverse collections of digital content result in new problems, including multiple communities of users, complex relationships among the digital objects, compound digital objects, shared behaviors, and so on. In recent years there has been considerable research and development of digital object repositories; however, they are in their early stages of development and most in the community have little experience with them. Current repositories utilize different hardware and software platforms, differing operating procedures and strategies, varied ingest mechanisms, and they use varying schemas and vocabularies.

The repository testbed will attempt to set up and evaluate instances of the major open source and some commercial digital object repositories. This testbed will provide a platform for the data collected as part of the overall project, but will also allow the project to investigate and document their strengths and weaknesses, their interoperability with one another, and their compliance to standards such as OAIS[2] and METS[3]. The repositories being evaluated include Fedora, DSpace, Greenstone, ePrints, and the OCLC Digital Archive.

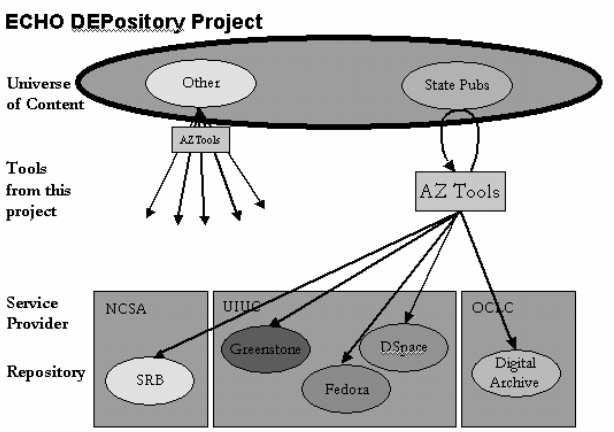

Figure 1. The ECHO DEPository Project

Based on this evaluation, publications and documentation will be created that will assist organizations in answering:

- What digital repository software is the most suitable for the storage and management of an institution's digital assets?
- What types of automatic harvesting and selection methods will work with various digital repository software systems?
- What does it take to move information from one repository into another, and how much of the original content and encoding structure can be preserved in that migration?

As part of the testbed, researchers will work with the different digital repositories and sample content on long-range research activities including techniques for migrating the semantic content of documents and document structures across future generations of encoding schemes – in other words, research into issues surrounding the long-term semantic preservation of digital resources.

Conclusion

Creating and evaluating the tools that will assist organizations in the ongoing process of curating – identifying, selecting, acquiring, managing, describing, and providing access to – their collections is vital if the community is to successfully ensure the preservation and continuing access to digital resources. With few tools available and few organizations which have experience with them, providing improved tools, such as the Web Archives Workbench, and practical advice based on real implement-

tations of digital repositories, the ECHO DEPository project will provide practical assistance for organizations working in the realm of digital preservation and access. This practical assistance will be supported and influenced by the research into long-term semantic preservation. The overall purpose of the project is to provide tools and information that are important to digital preservation currently, while undertaking the continuing research into the challenges of that preservation.

References

1. Pearce-Moses, Richard. A Glossary of Archival and Records Terminology (Society of American Archivists, forthcoming early 2005). Available from www.archivists.org/glossary/ [accessed 30 January 2005].
2. Open Archives Information System Reference Model, ISO 14721:2003.
3. Metadata Encoding Transmission Standard (METS). Available from http://www.loc.gov/standards/mets/ [accessed 30 January 2005]

Biographies

Judith Cobb is the Senior Product Specialist for the OCLC Digital Archive. The OCLC Digital Archive is an OAIS compliant repository that supports the management and preservation of digital materials. Judith holds a Bachelor's degree from Houghton College and a Masters of Library Science from the University of New York. She has 15 years experience working in the archives and library fields, specializing in the various aspects of digital preservation and access.

Richard Pearce-Moses is currently working as the Director of Digital Government Information for the Arizona State Library and Archives. He has a Master of Arts in American Studies from the University of Texas at Austin and a Master of Science in Library and Information Science from the University of Illinois at Urbana-Champaign. Pearce-Moses received a National Historical Publications and Records Commission Archival Research Fellowship to create a new glossary of archival and records terminology and is currently Vice President of the Society of American Archivists.

Taylor Surface is Global Product Manager, Digital Archiving Services, at OCLC. His team is developing OCLC's Digital Archive capability and other tools and services supporting cultural heritage institution's digital collection life cycle. Taylor has over 16 years of experience in information technology ranging from information retrieval research to development of information management services. Taylor's education includes a B.S. Computer Science / Engineering from Ohio State University and an M.B.A. from Capital University in Columbus, Ohio.