MetaScholar Initiative

General Libraries
540 Asbury Circle
Emory University
Atlanta, GA 30322

Phone 404 727 2204
Fax 404 727 0827

Workshop on Applications of Metadata Harvesting in Scholarly Portals

Findings from the MetaScholar Projects: AmericanSouth and MetaArchive

Edited by Martin Halbert

Emory University, Atlanta, GA

October 24, 2003

URL to this document: metascholar.org/documents/proceedings/MetaScholarFindingsProceedings.pdf

Table of Contents

Preface

It is with great pleasure that I preface these proceedings for the *Workshop on Applications of Metadata Harvesting in Scholarly Portals: Findings from the MetaScholar Projects: AmericanSouth and MetaArchive.* This event marks the conclusion of two projects funded by the Andrew W. Mellon Foundation to advance the understanding of metadata harvesting and scholarly communication. These projects were conjoined in 2001 to form the MetaScholar Initiative, an ongoing research collaboration based at Emory University involving librarians, scholars, archivists, curators, and technologists. Now concluding, these projects have successfully led to many findings that have informed the planning for new endeavors that the MetaScholar Initiative is now undertaking.

Our goals in organizing this workshop have been: to include presentations of research findings by participants in the projects; to provide a forum for scholars involved in the projects to discuss development of the portals for scholarly communication and research; to discuss future development of the resulting portal systems; to introduce new MetaScholar Projects now underway; and to examine the value of the OAI and other open source systems in the academic community.

I would like to thank the many people, institutions, and organizations that have made the MetaScholar Initiative possible. Enormous thanks go to the many partner libraries and archives that have been a part of this initiative. While the many individuals who have collaborated in these projects are far too numerous to list by name, I want to express my gratitude and thanks for the collegial relations we have enjoyed and all the work they have done in the course of the AmericanSouth and MetaArchive projects. These were investigative projects, and we were often exploring new territories, with all the confusion and bewilderment that such explorations can entail. Everyone has without exception been pleasant to work with, engaged in the process, and tolerant of the research process. I want to particularly thank all those who were willing to share their thoughts at this workshop, because the broad participation by many project partners makes this an invaluable opportunity to reflect back on the work we have accomplished and the knowledge we have gained.

I want to thank the leaders of the many partner institutions of the MetaScholar Initiative, particularly Kate Nevins of SOLINET, who had faith in our ability to carry out the goals of the AmericanSouth project, and of course Joan Gotwals (emeritus Vice-Provost and Director of Libraries at Emory University) and her successor Linda Matthews, who have been so supportive of these projects. Many thanks go to John Burger of ASERL and Sandy Nyberg of SOLINET, colleagues who have worked diligently in support of the AmericanSouth project.

Finally, great honor and thanks must be directed to Donald J. Waters and the other leaders of the Andrew W. Mellon Foundation. In supporting these and many other innovative projects over the years, the Mellon Foundation has led the way for the transformation of scholarly communication in the 21st century.

Martin Halbert
Emory University
October 23, 2003

Program Agenda

08:30 - 09:00 Registration and Coffee / Light Breakfast (provided)

The first half-hour is intended to provide attendees with time for informal chats with the central project staff, the project partner institution representatives, and the scholarly design team.

09:00 - 09:55 KEYNOTE PANEL

Don Waters (Program Officer, Andrew W. Mellon Foundation)
Martin Halbert (Director for Library Systems, Emory University)
Charles R. Wilson (Director, Center for Study of Southern Culture, University of Mississippi)

This session will review the initial research questions raised within the MetaScholar Initiative, and will explore the findings of the MetaScholar Initiative projects, AmericanSouth.Org and MetaArchive.Org, regarding applications of metadata harvesting to online scholarly communication portals

10:00 - 10:55 SESSION ONE: METADATA—OAI, DC and EAD

Sheila McAlister (Project Manager and Digital Metadata Coordinator, Digital Library of Georgia)
Susan Pinckard (Monocat Team Leader, Emory University)
Kathryn Stallard (Head, Special Collections & John G. Tower Archivist, Southwestern University)

This session will explore some of the issues raised for archivists and librarians by the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). What benefits and challenges do we encounter as we transform records into the Dublin Core format for OAI harvesting? How can we best disseminate EAD records and other forms of archival metadata through the OAI-PMH?

11:00 - 12:25 SESSION TWO: TECHNOLOGY—HARVESTERS AND PROVIDERS

Martin Halbert (Director for Library Systems, Emory University)
Jason White (Senior Systems Analyst, Emory University)
Chris Hodge (SunSITE Coordinator, Client Services, University of Tennessee, Knoxville)
Anthony Smith (Digital Initiatives Librarian, University of Tennessee, Knoxville)
Kurt Maly (Kaufman Professor and Chair, Department of Computer Science, Old Dominion University)

This session will discuss several key issues for technologists, especially concerning the best way to produce, test, and circulate the software tools needed for both metadata harvesting *and* for creating and sustaining a scholarly research portal. Technologies deployed in the MetaScholar projects will be reviewed.

12:30 - 01:25 Lunch (provided)

01:30 - 02:25 **SESSION THREE: COLLABORATION**

Linda Matthews (Vice Provost and Director of Libraries, Emory University)
Sheri Downer (Interim Dean, Auburn University Libraries)
Gordon Jones (Director of Exhibitions and Collections, Atlanta History Center)
Charles R. Wilson (Director, Center for Study of Southern Culture, University of Mississippi)

This session will consider the benefits of collaborative efforts that involve scholars, librarians, archivists, and technologists. What are the advantages of undertaking such collaborative national projects as AmericanSouth in order to create discovery services, share metadata, and publish work online via new technologies?

02:30 - 03:25 **SESSION FOUR: ONLINE SCHOLARSHIP**

Lucinda MacKethan (Professor of English, North Carolina State University)
Carole Merritt (Director, Herndon Home)
Erich Kesse (Director, Digital Library Center, George A. Smathers Libraries, University of Florida)
William Thomas (Professor & Director of the Center for Digital History, Univ. of Virginia)

This session will explore the new forms of scholarly expression made possible by the Web and realized through such research portals as AmericanSouth. Panelists will consider how digital media differs from print media for scholarship presentation, and what this may mean for the future of academic writings. They will also discuss the opportunities the Web provides for different research communities (i.e., comprised of both academic and non-academic researchers) to access and contribute to scholarly materials.

03:30 - 04:30 **SESSION FIVE: FUTURE STEPS**

Martin Halbert (Director for Library Systems, Emory University)
Katherine Skinner (MOSC Project Manager and Woodruff Fellow, Emory University)

This session will raise questions about metadata standards, regional cooperative efforts, and collaborative scholarly projects. Panelists will highlight several new OAI-PMH based projects that are beginning this year at Emory University, and will share their ideas about facilitating the use of the OAI-PMH and other technological standards among both large and small archives.

Findings from the MetaArchive and AmericanSouth Projects on the Application of Metadata Harvesting to Scholarly Portals

Martin Halbert, Emory University.

Summary

The MetaScholar Initiative has undertaken two projects concerning the application of metadata harvesting to scholarly communication: MetaArchive and AmericanSouth. The findings of these projects are briefly reviewed, as well as future plans.

Introduction

This article summarizes major findings of the MetaArchive and AmericanSouth projects, two of the seven projects of the 2001 Mellon Metadata Harvesting Initiative. This article will not deal with the activities undertaken in these projects, or the decision to conjoin the two projects to form the MetaScholar Initiative, as this information has been reported in detail elsewhere. [Halbert, 2003] Nor will it recount the already public motivations of the Mellon Foundation in funding these and other projects. [Waters, 2001] What will be reported are the general motivations that led to undertaking the MetaArchive and AmericanSouth projects, the questions that the projects set out to answer, and what findings were reached, both expected and unexpected.

Importance of Metadata Harvesting and Scholarly Communication

In the year 2000, I came to strongly believe that the emerging Open Archives Initiative Protocol for Metadata Harvesting was an important new technology that would have very significant impacts on the work of librarians and others who work to support learning communities. The development of the OAI-PMH was seen very quickly by many as holding great promise for various functions in digital repositories. [Lagoze, 2003] I felt that it was important for the Emory University General Libraries to quickly engage with other institutions in substantive efforts to experiment with the protocol to better understand the possibilities that it offered for transformative services that digital library endeavors could offer in support of scholarly communication. This area was clearly important for several reasons relating to the changing expectations of library users.

Many librarians have become concerned that commodified online services like Google are increasingly replacing libraries as the research tools of choice for many categories of library users, notably undergraduates, but increasingly graduate students and faculty. The power of search engines is remarkable, and can indeed produce remarkable results. But such commercial services are inherently subject to commercial bias, and generally index a large number of sites that are not sources of information that are credible and trustworthy

for academic inquiry. Clearly, if libraries are to remain relevant to learning communities, online discovery services offered by libraries and other repositories of scholarly information must evolve very rapidly to keep pace with systems like Google.

To accomplish this, I feel that libraries must recover the basic claim to scholarly relevance: providing organized access to the information that scholars need. Librarians have many talents to apply to this task. They deeply appreciate the value of information abstractions (indexes, catalogs, bibliographies) for providing organized access to information. Indeed, much of the library's strength as a culture is vested in metadata praxis built on centuries of experience with print materials.

It appears to me that we are at a critical cusp of librarianship, when we have the opportunity to build on our historical strengths in order to adapt to changing times. Gaining knowledge of how to manage metadata in the digital environment for scholarly communication purposes is key to the future of libraries, and it was essential for the library field to begin earnest endeavors engaging with this topic.

In funding a series of projects to explore this new territory, the Mellon Foundation moved the field forward significantly. The collaboration between Emory University and SOLINET that led to the conjunction of the AmericanSouth and MetaArchive projects is one of a large number of projects investigating the issues raised by the OAI-PMH for libraries. And yet the resulting MetaScholar Initiative based at Emory University has some unique areas of focus.

Main Research Areas of the MetaScholar Initiative

Our projects were described as feasibility studies for a large number of issues related to metadata harvesting and scholarly communication. The research of both projects essentially focuses on two broad areas:

1. How can a centralized group of staff and systems facilitate and catalyze widespread distribution and use of metadata about research collections?

2. How can scholars, librarians, and archivists engage in such a process to create effective tools for research and teaching?

Our projects quickly began to focus on topics much broader than metadata harvesting, but since these two areas of inquiry were built on the aspirations of the OAI-PMH, it is worth briefly reviewing some key points about the protocol.

Aims of the OAI

The Open Archives Initiative began in a 1999 meeting held in Santa Fe that was convened by various groups to seek ways of enhancing access to e-prints and other digital archives of value to scholars. Some of the key sponsors and participants in the OAI have been the Digital Library Federation (DLF), The Coalition for Networked Information (CNI), the National Science Foundation (NSF), and the Mellon Foundation.

The OAI articulated a general aim of developing and promoting interoperability standards to facilitate the efficient dissemination of metadata and content. To date, the Protocol for Metadata Harvesting has been the most prominent product of the OAI, although it has served as a valuable forum for metadata-related discussions.

The OAI strategy for sharing metadata was to create a protocol that:

- Could be simple to program on top of existing systems.
- Would enable digital repositories to provide bulk transfers of metadata concerning items held.
- Decoupled the task of searching metadata from storing metadata.
- Facilitated the creation of separate metadata searching services that will scale.

There were many hopes for the OAI-PMH in 1999, that amounted to big untested questions associated with the strategic aims of the protocol. Could it really enable automated distribution of metadata? Would the protocol in fact be simple to graft onto existing digital repositories? Could discovery systems be created by harvesting metadata? Would the protocol benefit scholarly communication?

Foci of the Two Projects

MetaArchive and AmericanSouth were basically designed as experimental tests of the strategic aims of the OAI-PMH, for the applied purpose of creating portals for scholarly communication applied to particular subject domains.

MetaArchive focused on smaller institutions (such as the archives of four-year liberal arts colleges, and small museums), as well as metadata aggregation services and techniques. Small institutions have very little IT infrastructure and are especially hard pressed to adopt new technologies in a timely way. Could a central staff of programmers and librarians deploy the OAI-PMH relatively quickly on behalf of smaller institutions to organize and disseminate information about their collections for scholarly use? What were affordable and effective software tools that could be applied to these questions? Could open source software solely comprise such an infrastructure? If the answers to these questions were positive, useful models for metadata harvesting in support of scholarly communication would be the result.

AmericanSouth focused on larger institutions (members of the Association of Southeastern Research Libraries, or ASERL), and intended to closely examine the questions of scholarly communication in a broad subject domain. How could online publishing services built up in association with metadata harvesting form an effective portal for scholarly communication?

We have come to some conclusions regarding these questions, summarized in the following sections.

Utility of the OAI-PMH

Can OAI-PMH capabilities be easily added to most digital repositories, and does this enable automated metadata dissemination? Our projects have accomplished many such implementations on a wide variety of infrastructures in the course of these 2 projects, enabling broad metadata dissemination. This general conclusion is now perhaps not very controversial, as many other infrastructures have used the protocol successfully for this purpose. What we can more specifically confirm is that metadata harvesting networks can be rapidly built at low cost using only open source tools.

Catalytic Actions to Foster OAI-PMH Adoption

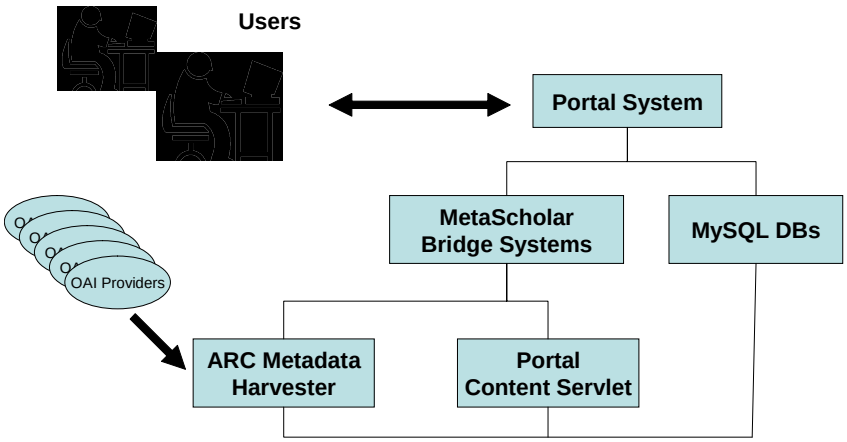
Can a small group of trained specialists effectively facilitate adoption of the protocol over a large region? Our project staff assisted in implementing many OAI providers in institutions ranging from New Jersey to Texas. Our interactions took many forms, ranging from simple consultation and advice for large organizations with their own IT staff, to efforts that were soup-to-nuts implementations for small institutions. We feel that such services are very beneficial to foster the widespread adoption of the protocol in order to disseminate collection metadata. As a result, we intend to develop a variety of metadata consultation services that can be offered to smaller libraries, archives, and museums at low cost (see section on next steps).

Utility of Open Source Software (OSS)

Software collaboratively developed by the community was more affordable and effective for our purposes than commercial software, at least for our projects. OSS has been considered a controversial strategy in the past, but is now increasingly embraced as a solid and reliable category of information technology by academic and even industry leaders. Emory University has a strategic commitment to using OSS in not only the MetaScholar Initiative, but throughout our digital library research infrastructure. OSS has consistently functioned better than any commercial software for new digital library projects.

The quality of the OSS we worked with from Virginia Tech and Old Dominion was outstanding. We could not have achieved our aims with closed commercial software because commercial vendors would have been unlikely to share source code of their products or (especially) allow others to modify such code. This was essential in adapting different systems to work together as a whole. The Open Digital Library (ODL) tools developed by Hussein Suleman and Rohit Kelapure (both then at Virginia Tech) were particularly adaptable to many OAI-PMH needs. [Suleman, 2001] The ARC software from Old Dominion is a powerful and comprehensive solution for metadata harvesting and indexing. [Liu, 2001] We were able to work with the digital library research groups at Virginia Tech and Old Dominion to adapt their software to our portal needs. By creating a cluster of bridge systems in Java as middleware, we effectively linked various components into a functioning whole.

Fig. 1: MetaScholar Technical Infrastructure



Metadata Aggregation & Searching

It has not proved easy to aggregate metadata using OAI providers and make it searchable. The metadata that can be gathered through harvesting is extremely heterogeneous, being produced in many different contexts. This phenomenon has been encountered by virtually all metadata harvesting projects, and has been described in detail elsewhere. [Halbert, ECDL 2003] While DC is indeed a Rosetta Stone for the protocol, it creates Procrustean dilemmas in metadata normalization that must be dealt with to provide effective searching, particularly in the case of Encoded Archival Description (EAD) metadata.

OAI-PMH Adoption

Despite work in our projects and many others, the majority of libraries and archives have not yet adopted the OAI protocol, as of 2003. Libraries and archives have been slow to adopt the protocol, because of a lack of technical staff and poor understanding of what the protocol is all about. They are, however, becoming increasingly interested, as evidenced by the regular inclusion of the topic in conference programs and publications. What this means is that many harvesting services (especially those for archival materials) will remain incomplete for some time to come, as key collections remain hidden. This is not a new phenomenon, in fact it has been the case for many years before the advent of the OAI-PMH that many special collections remained unknown and inaccessible. [Jones, 2003] The basic fact is that special collections are better funded for acquiring materials than processing them and making them accessible, whether through the OAI-PMH or other means.

Metadata Culture Clashes

Various groups have quite different assumptions about metadata and its functions, based on print applications. Catalogers and archivists embody such a culture clash, which plays out here as a mismatch between DC and EAD. Neither group is right or wrong, simply different in their expectations. Other workshop papers will discuss this issue in more detail. At the simplest level this is a difference in outlook concerning the best way to describe a body of material. Archivists are accustomed to preparing a single document describing an entire collection and subsequently adding as much detail to this document as possible. Catalogers and indexers are accustomed to describing a large and heterogeneous collection of books or records that have relatively uniform granularity. Metadata arising from the archival and cataloging perspectives is immiscible without additional processing. Put another way, it is jarring cognitively for a user to simultaneously retrieve both an extended EAD finding aid and a one line item record in a search. Another dichotomy arises when librarians come together with museum curators. Curators have yet another perspective relating to exhibitions and asset management.

Users of metadata also have quite different assumptions about metadata and what they want to be able to do with it. Focus groups demonstrated to us that students have very different expectations of what metadata a portal would provide when compared to the expectations of a faculty member. Students are more interested in prescriptive metadata that will guide them to successful completion of learning tasks (i.e. this is a good archival item for study). Faculty and advanced graduate scholars are preferentially interested in as much descriptive metadata as possible, and are in fact suspicious of prescriptive metadata. Academic learners are interested in academically generated metadata appropriate to their community, whereas public learners are attracted to metadata that reflect their proclivities and special interest groups. Finally, subject disciplines often have

quite specific organizing frameworks that they assume will be present in information. Creating portals can lead to 3-way metadata culture clashes, for example when trying to accommodate the expectations of students, faculty, and librarians. Trying to mediate such discussions can become very complicated. What is clear to us is that developing effective "scholarly portals" is going to require long term developmental relationships between varieties of involved groups working together collaboratively over time to overcome metadata culture clashes.

Importance of Preserving Context

Metadata is not context free. Most metadata harvesting projects have bumped up against the fact that metadata is always created in a well-constrained context. Take it out of that context and serious problems are likely to arise, especially when mixing it with other de-contextualized metadata. Harvesting efforts have to address themselves to preservation of context whenever possible. This can be done in a variety of ways through hyperlinks back to organizing frameworks, top level information, or cross-references. Annotations also allow an ad hoc capability for users to add context as they see the need arise.

Functions of Metadata

Metadata is not just descriptive in function. The DC mandate in the OAI-PMH emphasizes descriptive metadata, rather than administrative or structural metadata. This may be fine in a homogenous realm of atomic information objects, but the digital world is not that way. In a realm of nested and recombinant learning object glop, we will need to share much more sophisticated metadata in standard ways.

Scholarly Communication

The most basic question of these projects was whether or not applications of metadata harvesting are useful for purposes of scholarly communication. Our tentative answer is yes, but much more research is obviously needed to understand exactly how to best make use of these applications in support of scholarly research and teaching activities through portals. So far, our team of scholars, librarians, and technologists has found aggregated metadata useful for scholarly communication in several core activities:

1. **Discovery of previously unknown research items.** This is particularly useful for searching across repositories that are widely separated by physical distances. Many of our focus group participants were scholars that had to undertake long journeys just to get a general sense of what was in special collections, and felt our portal approach would pull together many relevant bodies of material for discovery purposes. This is the fundamental utility of traditional library catalogs, and the OAI-PMH does indeed enable automating generated catalogs of distributed holdings (albeit of very uneven metadata).
2. **Exposure of the hidden web.** This was one of the key goals of the Mellon Metadata Harvesting Initiative, and the protocol does enable exposure of information valuable for scholarly purposes and otherwise concealed in local database infrastructures.
3. **Enables the creation of an authoritative resource for searching.** Harvesting metadata selectively from archives and other repositories chosen by scholars and librarians creates an information resource that scholars have a clearer understanding of and more confidence in than commercial search engines.
4. **Provides a means of assigning emergent contexts.** Because the meaning of any particular research collection item is dependant on the perspective a scholar is coming from (see comments below on preserving context), agile mechanisms for

distributing, tracking, and enhancing metadata describing items is critical for creating emergent contexts for study. At the simplest level, providing an annotation capability for portal users has been found to be important by many harvesting services.

We intend to continue investigating these issues in new projects that will build on the relationships we have established in these initial activities. There are obviously many other projects underway that are studying the questions involved in developing scholarly portals. What differentiates our ongoing efforts will hopefully be the characteristics of our first two projects: the fact that they were close collaborations, used open source tools that can easily be adapted in new settings, and that they build on the interest in advancing scholarly communication held by Emory University and our other partner libraries.

The Need for Metadata Gardening

We found that there is still inadequate understanding of the entire cycle of metadata production, dissemination, reprocessing, and extended uses. Providing effective metadata services for scholarly communication will necessitate metadata "gardening," or efforts to effectively manage the entire cycle of metadata production and consumption. Without better perspective on how to coordinate this cycle of cultivating metadata providers, harvesting, and finally organizing metadata for discovery, we feel that coordination problems will continue to compromise the effectiveness of services based on metadata harvesting. This is very complex when considering the range of metadata sources potentially active on the Internet, and will require much more study to understand. We have previously done this only in smaller, closed communities (such as libraries cataloging their collections).

Metadata and the Web

There was a hope that metadata harvesting would by itself provide a strong alternative to web search engines. Many harvesting experts and projects (MetaScholar, ARC, Herbert Von De Sompel) have now concluded that some combination of metadata and web indexing must be explored. We intend to undertake a new project to experiment with exactly this area.

Conclusions and Next Steps

We have learned a few things in the last few years about metadata and scholarly communication, but we have also started learning how much more we need to learn.
The future is bright with new possibilities for collaborative endeavors to advance scholarly communication. The MetaScholar Initiative will undertake the following projects in coming months:

- Fostering adoption of the OAI-PMH through workshops such as this one.
- Developing a variety of metadata consultation services that can be offered to smaller libraries, archives, and museums at low cost.
- Continuing to develop the AmericanSouth portal in collaboration with the Center for the Study of Southern Culture, SOLINET, and other groups.
- Clarifying the relationship of metadata harvesting services and other discovery services through our new MetaCombine project, to explore the combination of metadata harvesting and web crawling, as well as semantic clustering of harvested metadata.

Bibliography

Cole, 2003 — Timothy W. Cole. "Using OAI: Innovations in the Sharing of Information." *Library Hi Tech*, Vol. 21, No. 2 (2003): 115-117.

Halbert, 2003 — Martin Halbert. "The MetaScholar Initiative: AmericanSouth.Org and MetaArchive.Org." *Library Hi Tech*, Vol. 21, No. 2 (2003): 182-198.

Halbert, ECDL 2003 — Martin Halbert, Joanne Kaczmarek, and Kat Hagedorn. "Findings from the Mellon Metadata Harvesting Initiative." *Research and Advanced Technology for Digital Libraries, Proceedings of the 7th European Conference, ECDL 2003,* Trondheim, Norway, Traugott Koch and Torvik Solvberg (eds.), Lecture Notes in Computer Science #2769, Springer Verlag: 58-69.

Jones, 2003 — Barbara M. Jones, et al. "Hidden Collections, Scholarly Barriers: Creating Access to Unprocessed Special Collections Materials in North America's Research Libraries." *A White Paper for the Association of Research Libraries Task Force on Special Collections.* Prepared for the *ARL Workshop on Exposing Hidden Collections*, September 8-9, 2003. Available online at URL: http://www.arl.org/collect/spcoll/ehc/HiddenCollsWhitePaperJun6.pdf

Lagoze, 2003 — Carl Lagoze and Herbert Van de Sompel. "The Making of the Open Archives Initiative Protocol for Metadata Harvesting." *Library Hi Tech*, Vol. 21, No. 2 (2003): 118-128.

Liu, 2001 — Xiaoming Liu, Kurt Maly, Mohammad Zubair, and Michael L. Nelson. "Arc - An OAI Service Provider for Digital Library Federation." *D-Lib Magazine*, Vol. 7, No. 4, April 2001. URL: http://www.dlib.org/dlib/april01/liu/04liu.html

Suleman, 2001 — Hussein Suleman and Ed Fox. "A Framework for Building Open Digital Libraries." *D-Lib Magazine*, Vol. 7, No. 12, December 2001. URL: http://www.dlib.org/dlib/december01/suleman/12suleman.html

Waters, 2001 — Donald J. Waters. "The Metadata Harvesting Initiative of the Mellon Foundation." *ARL Bimonthly Report* 217 (August 2001): 10-11.

A Scholar's Perspective on AmericanSouth.Org

Dr. Charles Reagan Wilson, Center for the Study of Southern Culture, University of Mississippi

Summary

An online portal has immense potential for organizing research and discussion of scholarly topics. Scholars of the history and culture of the American South have shaped the development of AmericanSouth.org, working with archivists, librarians, and technologists. The portal contains metadata from selected archives and special collections, guiding users of the site to rich information about the South.

The Scholarly Design Team has provided contextual material through extensive reference guides, shorter online articles, methodological guides, curriculum resources, and lists of authoritative Web sites. The scholarly focus has been on ideas of "region" and "place" as a way to place metadata and information in digitized archival collections in a broader intellectual perspective.

Introduction

When I first heard about AmericanSouth.org, I was excited about the potential to make use of advanced technology to create easier access to information about archival materials related to the study of the American South. Scholars working on AmericanSouth.org have adapted traditional models of scholarship to the new media we are using, in order to preserve scholarly standards while taking full advantage of the capabilities of Web-based research. From the beginning those of us working on the project had the goal of fostering new ways of communication among scholars, realizing that the portal had the potential to be more than a database of archival metadata. The basis of everything we have done, to be sure, is the primary sources that will provide information and connections to digitized archival materials, but we have attempted to begin the process of creating context for online research about the South.

AmericanSouth.org promises to advance dramatically research on the region, making available widely scattered information, organizing primary sources, and providing ways of using them that will make the individual primary sources part of a larger project.

Scholarly Design Tean

In selecting members of the Scholarly Design Team (SDT) who would think systematically about creating an online scholarly community around issues related to the American South, we sought scholars who had already produced notable work on the South. We wanted an interdisciplinary mix of researchers who could bring their academic specialties into a broader, collaborative discussion of the South. Finally, we looked for scholars who seemed attuned to exploring the capabilities of online services for scholarship.

I come at this project from the background of a historian interested in multidisciplinary studies, particularly the study of region and culture. As director of the Center for the Study of Southern Culture, I have worked to make use of Web based research, using the Center's Web site to showcase documentary studies fieldwork projects and to make accessible photographs, films, oral history, and business records relating to one particular plantation, Perthshire in the Mississippi Delta, allowing in-depth research using just materials on the Web site. We now are preparing the "Mississippi Encyclopedia," a one-volume reference book to appear in print and online.

Other members of the SDT bring different perspectives. Allen Tullos (Emory University) is also a historian but one with a special interest in the music of the South, having taught frequent courses making using of online curriculum materials. Lucinda MacKethan (North Carolina State University) is a literary scholar who is coeditor of the Companion to Southern Literature, a definitive and imaginative reference work that puts literature within a broader cultural framework. Carole Merritt (Herndon Home historic site, Atlanta) brings the perspective of public history, having published on a leading African American family in the South and on issues of the presentation of historical houses to a broad public. Will Thomas, the director of the Virginia Center for Digital History (University of Virginia), was involved with "The Valley of the Shadow," a landmark multimedia project on the American Civil War, and he has published extensively about online historical research.

Adapting Scholarly Models, Creating New Ones

One traditional scholarly model that has influenced AmericanSouth.org is the reference work, in this case the Encyclopedia of Southern Culture. The Encyclopedia charted the cultural landscape of the South, drawing from such disparate disciplines as history, geography, folklore, linguistics, political science, literary studies, and sociology. Thematic areas ranged from religion to politics, social class to mythology, violence to social manners. It was a truly collaborative project, with over 700 contributors. A reference book is indeed one good model for contextual material online, bringing together scholars to give authoritative, factual, and well-organized material to users. They synthesize existing scholarship and provide their information through a balanced, objective tone. Achieving these characteristics will give AmericanSouth.org a credibility associated with solid, multidisciplinary scholarship. We will not only make use of the reference work model in general, but we are literally working with the University of North Carolina Press to work out the details of putting the Encyclopedia of Southern Culture online, to provide a systematic contextual framework that will be further enriched by other articles prepared by the Scholarly Design Team.

We anticipate in the future adapting another scholarly model—the journal. We have already begun approaching conference organizers with ideas of having refereed evaluations of papers that will appear on the Web site. Thanks to Lucinda MacKethan we should be able to do this soon with selected papers from the Southern Writers Symposium. We are also commissioning a series of specialized articles on regions in the South, to supplement the original articles now online written by the Scholarly Design Team.

We are also creating new scholarly models, drawing from the ability of the portal to organize existing materials on the Web. We are making use of existing commentaries on recent research about the South, identifying and making accessible discussions of archival collection materials that are part of the project. We hope to expand this feature to include reviews of monographs that have used AmericanSouth.org collections. We are providing guides to methodologies that have been used to study the South, and in the process we hope to point users of the site to differing ways that material can be analyzed and interpreted.

We have begun identifying reliable Web sites that contain useful information about the South, and we want to play a constructive role as something of a gatekeeper in pointing users to sites that have authoritative and significant information about the South. While this had not been an original goal of the SDT, we soon realized that playing this role would be a valuable contribution to those trying to find reliable information online about the region.

Communication

AmericanSouth.org will include an important component of online dialogue—interactive communication among users of the portal. In addition to the reference guies, online articles, methodological guides, curriculum resources, and authoritative Web sites, the portal will provide the opportunity for users to add commentary to advance discussion of key topics related to the history and culture of the American South.

Members of the SDT themselves engaged in such dialogue, commenting on each other's articles when they appeared on the preliminary Web site. These discussions typically focused on issues of region and place, often leading to theoretical commentaries. In considering how many regions the South contains, Allen Tullos suggested that we not limit our list to an arbitrary number but provide an expansive group of essays on regions, which contributors to the Web site in the future can shape. Will Thomas agreed with this, noting that "we should try to take full advantage of the openness that the technology affords us." That has been our goal throughout, as we have expanded features to enable scholars to grow the site.

In all of our work we have tried to be multidisciplinary, making scholars in different fields aware of the usefulness of archival and special collection materials to address a variety of topics and organizing existing online materials.

Regions and Places

The work of the SDT is revealing of the dialogue that will likely proceed in nurturing an online scholarly community. In our planning work, members of the SDT discussed numerous thematic approaches to organizing our work, but we came to agree that region and place provided the best focus. The idea of "the South" is a historical term, a cultural construction with deep mythical resonances that continue to be exploited for commercial and other purposes. These clearly need to be studied and understood. Much contemporary work on region builds on theoretical work by geographers, historians, anthropologists, folklorists, and others that see "place" as the defining concept for understanding regions, and the SDT has used "place" as a useful organizing category.

Each scholar has written a reference guide to a major topic, such as religion or music. Lucinda MacKethan provided an original overview of southern literature by looking at it through the concept of genre. Carole Merritt was more focused, studying "African American Community Building in Atlanta," which was a broader guide to studying race in the United States. Will Thomas has written on "The Segregation of Information: Media, History, and Civil Rights in the Twentieth Century South," which will include links to media resources. Allen Tullos has made impressive use of online musical resources in writing guides to various genres of the music of the South. I have written on religion in the South, a historical and cultural narrative of important themes and with a historiographical assessment of existing secondary literature and links to a variety of Web sites.

Conclusion

In all of our work, we have tried to be multidisciplinary, making scholars in different fields aware of the usefulness of archival and special collections materials to address a variety of topics and organizing existing online resources. As scholars we have been particularly attuned to the scholarly audience and ways to combine solid scholarly standards while taking advantage of the unique opportunities for presentation of information and dialogue about it that the Web presents. We have also tried to avoid jargon and write clearly in order to make our materials accessible to students and a broad general audience. We are in a process of discovery, and in the future we hope to define new features for the portal that will make clearer connections between our contextual material on the South and the specific information contained in the archival collections to which metadata on the site will give increasing access.

Group Think in a World of Individuals: OAI, Metadata, and Sharing Cultural Resources

Sheila McAlister, Digital Library of Georgia, University of Georgia

Summary

The Open Archives Initiative Metadata Harvesting Protocol (OAI) supports the potential for sharing metadata across cultural heritage institutions. The creation of large repositories of such information is advantageous to scholars and others as it brings them closer to one-stop-shopping for cultural resources. However, the mixing of metadata from a variety of institutions and communities poses difficulties for discovery and interoperability. The paper discusses some of the barriers to interoperability, examines archival description in the online environment and in OAI repositories, and concludes by examining future cooperative metadata efforts at the Digital Library of Georgia and the emerging archival context framework Encoded Archival Context.

OAI: A Great Melting Pot

Emerging out of the e-print community, OAI was designed to allow owners of metadata sets simple ways of exposing their data for harvesting by others. Metadata aggregators could then provide access to information about research materials that are geographically dispersed and in different types of institutions. The gathered data could be repurposed or enhanced by the collecting agency as is the case with AmericanSouth's scholarly portal services. Not only are users exposed to a wide range of materials treating the history and culture of the South, they can also find scholarly articles, tools for interpretation, and avenues for scholarly discussion. While single point access to a variety of metadata for archival and other special collections materials is of benefit to both scholars and repositories alike, metadata practices at the contributing institutions may hinder discovery.

The materials found within cultural heritage institutions encompass the well-known as well as the unknown. Often the materials with common creators, provenance, and subjects are dispersed across multiple institutions. If each institution uses an ambiguous form of the name, retrieval may be hampered as searchers are unable to distinguish between John Smith and John Smith. Authority control is a time consuming and expensive venture because of the research involved in creating, analyzing, and maintaining names, but without it interoperability is hampered.

A variety of community-specific thesauri exist for describing cultural heritage materials. Within the library community both Library of Congress Subject Heading and Sears Subjects are employed. The archivists may employ LCSH as well as terms from the Art and Architecture Thesaurus and the Thesaurus for Graphic Materials. Discipline-specific indexing terms may also be employed if metadata records are created by a subject specialist as is the case with records from the Florissant Fossil Beds National Monument found in the Colorado Digital Library's shared metadata catalog. If seamless searching is

to be attained, these knowledge organization systems and subject vocabularies must be made interoperable. International efforts thus far have focused on mapping and integration of such vocabularies or on the creation of new organization systems (Chan and Zeng, 2002, 2).

Depending on the scope of institutions contributing records to a repository, there may be several different approaches to description. For example, within the library community traditionally all records must have a title. The museum community, by contrast, does not generally create titles for items. Even within the same type of organization, adoption of standards differs. In discussing the Colorado Digital Library's efforts to create a catalog of metadata for a variety of cultural heritage institutions, Liz Bishoff noted such differences among museums, "Within the museum community, specific segments have standards, such as art museums; however across the museum community there are few commonly adopted standards." (Bishoff, 2000, 3). To combat such varying descriptive practices, the Mellon-funded OAI project at the University of Illinois Urbana Champaign analyzed approximately 613,813 metadata records from 23 of the metadata providers who contributed to the University of Illinois Cultural Heritage Repository to determine the variability of metadata practices across communities. In addition, the project team analyzed what elements were populated the most and least consistently and differences in content in DC elements such as date and coverage. The Illinois team suggests that metadata providers place their records into OAI sets according to subject and type of material and that portal services consider searching subject and description fields together and contributor and creator elements together (Shreeves, Kaczmarek, and Cole, 2003).

Archival Description in an Online Environment

As mentioned above, different cultural heritage communities have different outlooks on description. Archives are no exception. Unlike museum and library materials, the basic unit for archival description is based on groups rather than on discrete items. The center of archival description is the creator; archival description is *provenance*-based. Archivists contend that in order to fully understand a creator, one must attempt to maintain the original order of its outputs (i.e., records or papers). For, how the individual or organization organized its files can provide illuminating evidence on its hows and whys. Because of the value of illuminating the relationships among a creator's records, archivists maintain *original order.* Within the library community, by contrast, emphasis is placed on collocation, like subjects together, rather than on exposing the organic relationships between materials. Because of the importance of these relationships and the often complex and varied nature of archival materials, archival description must elucidate not only content but also the context of creation. Emphasis is based on collective description and moving from description of the whole to description of subordinate parts.

Over the years archival descriptive practices and standards have changed. Over the past two decades with the advent of increased automation, new standards have evolved within the community to facilitate the sharing of descriptive information such as *Archives, Personal Papers and Manuscripts* (APPM), a content and format standard; *General International Standard Archival Description* (ISAD(G)), a data element standard; *International Standard Archival Authority Record for Corporate Bodies, Persons and Families* (ISAAR(CPF)), a data format standard for archival authority records; and USMARC and Encoded Archival Description (EAD), data structure standards. Despite this evolution, many descriptive tools within archives and special collections pre-date the re-examination of practices. As a result, many collections described in the past may not integrate well with newer description. For example, while some collections are calendared, others list only folder titles with no contextual or other descriptive information, and yet

others combine biographical and content information into a single essay. Many descriptions do not exist in digital form. Retrospective conversion is a time-consuming and expensive process and must be weighed against a repository's other priorities. Descriptive information for such collections may not even make it into OAI repositories such as AmericanSouth, because a repository may not have the resources to upgrade the description into digital form or it may not wish to share description it may now consider less than adequate.

USMARC (MARC AMC in particular), one of the first data formats used for the exchange of archival description, has been used widely as evidenced by cooperative archival cataloging projects such as the Georgia-based Georgia Archives and Manuscripts AutoMated Access (GAMMA) and Cooperative Historically Black Colleges and Universities Archival Survey Project (CHASP) projects. While USMARC is valuable as a means for integrating archival description in shared bibliographic utilities such as OCLC and RLIN and local library catalogs, the format is not ideal for archival description. By nature, MARC records are flat and do not support fully the expression of hierarchical relationships. Their limited size also forces catalogers to excise greatly the rich description normally found within a finding aid. Concise yet deep description is of utmost importance. Such brevity may mask pertinent materials or direct researchers to collections that may contain relevant items (Pitti, 1998, 11). The MARC record may be useful as a pointer to fuller description through the use of the "finding aid available" note, field 555, or the "electronic location and access" field, 856.

An outgrowth of the Berkeley Finding Aid Project, Encoded Archival Description (EAD) seeks to make the rich, multi-leveled description of finding aids available in an online environment through SGML/XML encoding. While MARC records can identify potential collections of use, the ability to search full descriptions of archival materials has the potential of allowing scholars and others to pinpoint germane materials. Developed within a community traditionally resistant to standardization, EAD was created to embrace the wide variety of descriptive practices.

> Despite a somewhat traditional penchant within the profession for the development of unique solutions to common problems (and a concomitant resistance to various types of standardization), many archivists instinctively are attracted to a technique that promises to reduce the needs to reinvent the finding aid wheel in every repository, or to rekey or edit data every time a software upgrade is necessary, and which also demonstrates clear potential to radically improve access to archival materials by facilitating structured access via the Internet (Dooley, 1998, 1).

Yet, EAD's flexibility is, in many ways, an impediment to interoperability. In a 2002 session at the Society of American Archivists annual meeting entitled "Challenges and Opportunities in Integrating Access," speakers Elizabeth Shaw, Clay Redding, and Christopher Prom highlighted the disparity in descriptive and encoding practices among repositories implementing EAD and the resulting difficulties in retrieval. The Encoded Archival Description Working Group shied away from providing data format and content standards in their *Applications Guidelines*. As of yet, there is no content standard for finding aids nor format standards for EAD. Best practices and guidelines have been developed by numerous state-wide EAD projects as well as by the Research Library Group. By continuing to work internationally to develop data content and format standards, the archival community may be able to take full advantage of the potential of EAD.

The Group vs. the Individual: EAD in the OAI Environment

Part of the promise of EAD is the ability to expose researchers to fuller descriptions of subordinate units of description and to maintain the contextual linkages provided by hierarchical description. How then does EAD play in the OAI environment which, at a minimum, provides access to flat, unqualified Dublin Core (DC) records? While the OAI protocol permits the provision of fuller, richer, associated metadata, the DC record may merely point to the encoded finding as in the MARC environment. In OAI repositories such as MetaScholar, one is confronted with the above-mentioned variation in practices. Some institutions are able to provide folder-level access to their collections through the unqualified DC records while others choose to provide only description akin to that found in MARC records yet less granular. The designed flexibility of DC may also impact the discovery of archival materials. While ISAD(G) specifies six required elements for archival description: reference code, title, creator, dates, extent of the described unit, and level of description, some of the file-level description found in OAI repositories contains neither the level of description nor the extent of the described unit. Will researchers be able to interpret the amount of potential information when they are unaware of whether the description found concerns a collection, a series, a folder, or an item? In splitting apart the EAD finding aid into multiple DC records, are researchers and archivists losing the context and hierarchy so important to the description and understanding of archival materials?

In the University of Illinois' Mellon-funded OAI project, archivist Chris Prom and other researchers tackled these two questions. When splitting apart encoded finding aids, the researchers at Illinois attempted to maintain references to an individual DC record's place in the hierarchy through use of a textual description in the relation field. The practice became problematic as it resulted in false hits and in extremely large record sizes. To balance the retrieval and file size issues with the need to preserve context, the project employed an XPOINTER back to the corresponding point in the finding aid provided outside of the search environment, but removed textual references to most levels of the hierarchy from the relation field (Prom, 2003). Nevertheless, Prom believes that OAI is promising as a method of exchange for archival description rather than as a description format. He posits that the use of the simple unqualified DC record employed by OAI may be used as a bridge between finding aids encoded in EAD following different markup styles. The simplified records derived from these EAD instances could be used for behind the scenes searching easing the problems inherent in non-standard encoding practices.

OAI and the Digital Library of Georgia

In thinking about future scholarly portals, standards, and regional cooperative efforts, I'd like to begin by considering my institution's own immediate plans for such undertakings. Over the next year, we, at the Digital Library of Georgia, will be studying closely the results and discoveries of the Mellon-funded OAI projects as we begin to develop our own metadata repository. By this winter, our collections will consist of over 70,000 digitized items including aerial photographs, photographs from the 19th and 20th centuries related to the architecture, landscapes, and people of the state, full-text searchable Georgia-related books, transcribed and annotated 19th century manuscripts and diaries, legislative and government documents, and 20th century editorial cartoons. Currently, our users must search each site or database to locate the materials they desire. After the launch of the repository, they will be able to search across all our resources, and, in the future, across the collections of other institutions with Georgia-related digital resources. Our records and those of contributing institutions will continue to be made available for harvesting by the AmericanSouth project.

As the Digital Library grows and expands its base of contributing institutions, many of the issues mentioned above will become increasingly acute. Our contributors will include not only academic institutions throughout the state but also public libraries, historical societies, and hopefully small museums. The materials found in these smaller, non-academic institutions, while rich in content, are often unknown to scholars and other researchers. In order to facilitate searching across our metadata repository and the EAD database we hope to launch, we will need to look at the issues of authority control and the use of multiple vocabularies. Already, we are developing a cross-project name authority database. Additionally, we are considering how to integrate ontologies or a thesaurus into our repository.

Using the metadata repository as a departure point, we hope next to create thematic approaches to our collections akin in many ways to the MetaScholar framework. While most of our projects already contain value-added resources such as bibliographies, lists of related archival materials, and essays, they are aimed at a more narrow audience, high school to undergraduate students and do not link across collections. Working with scholarly advisors and education specialists, we hope to create similar contextual tools to complement prominent themes found within our materials. As the Digital Library serves multiple audiences, we hope to create sites that will be adaptive. We may, for example, have separate resources for elementary, middle, and high school students as well as resources for teachers, hobbyists, and other life-long learners.

Encoded Archival Context

On a larger, international scale, Encoded Archival Context, an international initiative to design an XML-based framework for encoding descriptions of record creators, looks as if it will assist in disseminating not only authority data about creators but also biographical and administrative history information. Informed by the International Standard Archival Authority Records for Corporate Bodies, Person and Families (ISAAR(CPF)), the framework will facilitate the exchange and sharing of creator description and provide shared contextual information not only for archives, but also for libraries and museums. Such research is labor-intensive and expensive, and sharing these descriptions could be an economic boon for the cultural heritage community. The benefits of EAC are not confined only to libraries, archives, and museums. According to Daniel Pitti, a member of the international team developing the prototype standard, standardization of creator description could facilitate the creation of regional biographical databases, a benefit to researchers and others (Pitti, 2003).

Conclusion

The OAI Metadata Protocol provides opportunities not only for collecting related metadata and pointing researchers and others to materials they seek without requiring a time-intensive institution-by-institution search. Moreover, these large collections of descriptive information provide occasions for the creation additional pedagogic or research-related frameworks. While interoperability issues exist when mining diverse data sets, current and future research projects are already investigating possible solutions. The MetaScholar Initiative, for example, will be investigating issues of authority control and semantic linking over the next two years. The Illinois project provides an excellent groundwork for alleviating retrieval issues with cross-community records and points to new ways that archival description might harness OAI's exchange format. The OAIster project at the University of Michigan also provides end-user data that will be helpful as more institutions set up data repositories. These endeavors will certainly boost the usefulness of OAI.

References

Bishoff, L. (2000), “Interoperability and standards in a museum library collaborative: the Colorado Digitization Project,” *First Monday* Vol. 5, no. 6; available at http://firstmonday.org/issues/issue5_6/bishoff/index.html (accessed 17 October 2003).

Chan, L.M. and Zeng, M.L. (2002), “Ensuring interoperability among subject vocabularies and knowledge organization schemes: a methodological analysis,” paper presented at the 68th IFLA Council and General Conference, Glasgow, Scotland, August 18-24; available at http://www.ifla.org/IV/ifla68/papers/008-122e.pdf (accessed 17 October 2003).

Dooley, J.M. (1998), “Introduction,” in Dooley, J.M. (ed), *Encoded Archival Description: Context, Theory, and Case Studies*, Society of American Archivists, Chicago, IL, 1998, pp. 1-4.

Halbert, M. “The Meatascholar Initiative: AmericanSouth.org and MetaArchive.org,” *Library Hi Tech*, Vol. 21, no. 2, pp. 182-198.

Hensen, S.L. (1998), “‘NISTF II’ and EAD: The evolution of archival description,” in Dooley, J.M. (ed), *Encoded Archival Description: Context, Theory, and Case Studies*, Society of American Archivists, Chicago, IL, 1998, pp. 23-34.

ISAG (G): General International Standard Archival Description: Adopted by the Committee on Descriptive Standards, Stockholm, Sweden, 19-22 September 1999. 2nd ed. International Council on Archives, 2000.

Pitti, D.V. (1995), "The Berkeley Finding Aid Project: Standards in navigation," in *Filling the Pipeline and Paying the Piper*, Association of Research Libraries, Washington, DC, pp. 161-166.

Pitti, D.V. (1998), “Encoded Archival Description: The development of an encoding standard for archival finding aids,” in Dooley, J.M. (ed), *Encoded Archival Description: Context, Theory, and Case Studies*, Society of American Archivists, Chicago, IL, 1998, pp. 7-22.

Pitti, D.V. (2003), "Encoded Archival Context (EAC),” paper presented at the International Conference Authority Control, Florence, Italy February 10-12; available at : http://www.unifi.it/universita/biblioteche/ac/relazioni/pitti_eng.pdf (accessed 17 October 2003).

Prom, C.J. (2003), “Reengineering archival access through the OAI protocols,” *Library Hi Tech*, Vol. 21, no. 2, pp. 199-209.

Shreeves, S.L., Kaczmarek, J.S. and Cole, T.W. (2003), “Harvesting cultural heritage metadata using the OAI protocol,” *Library Hi Tech*, Vol. 21, no. 2, pp. 159-69.

Thoughts on Metadata Crosswalks and Harvesting

Susan H. Pinckard, Emory University

Summary

Metadata harvesting is about providing access to collections of information that, because their existence or contents may not be widely known, have remained on the fringe of scholarly research. Effective metadata crosswalks are a key element in the process of producing effective metadata for dissemination as DC through the OAI-PMH.

So what exactly is metadata harvesting and why is it important?

The usual definition states that metadata is data about data. It's really just information about information, structured to feed into automated processes (i.e. a metadata harvester or your online catalog), as a means to furthering resource discovery. By providing information-rich records (metadata) to a project like the MetaScholar Initiative, scholars have a greater likelihood of locating materials from **your** collection that will enrich their research.

Metadata crosswalks

Crosswalks are the means by which the fields or data elements in one metadata standard, such as MARC, Dublin Core, or EAD, can be semantically mapped to the fields or data elements of another standard. In other words, they convert data from one metadata standard to another by mapping fields with similar functions or meanings. The challenge in effectively utilizing crosswalks lies in the nature of the contributed metadata records themselves:

- Some records may be so data rich that mapping their fields to Dublin Core, our default standard, almost seems like "dumbing" down the information.
- Some records may be so brief that the resulting DC record is truly "bare bones."

Decisions on how to map similar fields led to in-depth discussions on the minimal level/amount of metadata needed from our partner institutions in order to successfully test the project.

Metadata Quality Control Issues

The usefulness of metadata is directly related to its quality. Long-standing cataloging practices such as controlled vocabularies, subject taxonomies, and authority control are best-practices for creating high quality metadata that can be used for browsing and other search functions. Metadata that is generated without implementing cataloging best practices suffers from poor usability. We found wide variation in the metadata received through our metadata harvests, and often realized that fixing the metadata after the fact by hand was simply not cost effective.

Key cataloging findings from MetaArchive Project

Access is the foundation for this and similar projects: metadata records should ideally address four functions. They should allow us to **find** (provide access points by which collections can be found), to **identify** (describe so as to enable correct interpretation), to **select** (provide means for selection of identified materials), and to **access** (explain how a scholar gains access to identified materials).

Access is the foundation for this and similar projects: in this case, more **is** usually better. Technical issues aside, without viable, reasonably full metadata records, collections will remain largely unfound by harvester users.

Access is the foundation for this and similar projects: there is a need for a cooperatively developed set of standards for metadata record creation.

.

MetaArchive: A Model for Facilitating Dissemination of Metadata from Small Archives

Kathryn Stallard, Southwestern University

Summary

Southwestern University library's Special Collections, like all special collections, holds unique materials available nowhere else. While collection-level MARC records can advertise some of these works to the universe of potential researchers, other works or collections present a challenge. This paper presents a case study of how the Mellon Foundation and Emory University helped a small institution launch into cyberspace a database of over 18,000 records for the Senator John G. Tower Papers.

The Collection: Senator John G. Tower Papers

John Tower, the first Republican elected to the Senate from Texas since Reconstruction, served from 1961 until 1985. His career spanned a fascinating and often troubled time in our country's history, and his papers mirror the period and provide insight into the events of the 1960s, 1970s, and 1980s. Covering the Vietnam War, civil rights, Watergate, women's issues, gun control, environmental concerns, the papers are a microcosm of the era from a national and Texas perspective, and scholars seek them out.

Tower's papers came to his alma mater in the mid-1980s, but a concentrated effort to process them started only after his death in 1991. One FTE professional archivist, one FTE support staff, and several student workers were hired for two years to appraise, arrange, and describe this very large and significant collection. A year or so after Tower's death, his family donated his extensive pre and post-Senate papers, and the team gained a year's extension. The project concluded with the completion of a printed general guide to the collection, a detailed descriptive guide for serious researchers, and a database that controlled the collection on the box and folder level.

Describe It and They Will Come

Three things determine real estate value: location, location, location. For rare books, it's condition, condition, condition. And for archival collections, it's access, access, access. If information can't be found, it has no value. The larger the collection, the more problematical is the ability to find information.

The decision as to how to provide access was among the first challenges facing the Tower project. We new that we would control the collection at the box and folder level, as is standard for most archival collections. We considered several relational database software packages, including MicroMarc, but we chose ProCite for several reasons: It

was inexpensive; search features were powerful yet simple and flexible; one of us knew it; and it has a gentle learning curve. The last factor was important for two reasons: we had little time to get up and running, and students would be doing a significant portion of the data entry. ProCite served us well—and continues to do so.

As we processed and refoldered the collection, we made note of important subjects and correspondents on the outside of those folders that contained "rich" information. By "rich" I mean substantive correspondence with an important figure, unique information about an important event, etc. An informal subject thesaurus, created as we worked, controlled some key terms, but we were also relying on ProCite's flexible search engine to help us locate materials researchers needed. During data entry, the names and subjects recorded on the outside of the folders were entered into ProCite's abstract field, thereby expanding the subject access to the collection.

In addition, once the collection was processed, we created a MARC record for the finding aid, which serves as a collection level record. We intended to create MARC records for the series level, but found that need never drove that intention sufficiently to hoist it to the top of the pile of the cataloging backlog.

If success is measured by researcher satisfaction, then we have triumphed: We receive glowing comments and letters of thanks from our users, who are effusive in their praise. Meanwhile, it's difficult not to wince as one sees a dependence on the database and an unwillingness to slog through materials that might have valuable information but didn't pull up in a database search. The database does not provide item level access, although some users seem to want to pretend it does. For example, a researcher studying Tower's courtship of Mexican-American voters was unwilling to go beyond those file folders that specifically pulled up on a search despite suggestions that he look at Tower's press office publications, voter analysis studies, and other files. The allure and lure of the database is so seductive, in fact, that it is the rare scholar who will even bother to read our award-winning *Guide to the John G. Tower Papers* much less the extensive descriptive guide written specifically for serious scholars.

Sharing Metadata: The Challenge

In the late 1990s we transformed the general printed guide to the Tower papers into a website. The guide contains a preface, biography, timeline, overview of the collection, and descriptions of the various series and subseries. Series and subseries information was listed at the box level, but did not extend to folder level information. Consequently, we purchased ProCite's web software in order to make the complete 18,000-plus record database available online. We were in the midst of wrestling with this software when the opportunity to participate in the MetaArchive project arose. Although the main ProCite database supports individual records with over 40 searchable fields, the web version supported only 5 fields. As a consequence, we were preparing to create a special database that would merge some data fields and eliminate others—it would not be a pretty or safe "crosswalk." We were also wondering if we should just stall, and hope that ProCite's parent company would upgrade the web software to match the sophistication of ProCite. We were hesitating at this stage when Martin Halbert invited Southwestern to participate in the MetaArchives project.

The Emory University Mellon-funded project appeared, *deus ex machina,* to solve the challenge of presenting our metadata to the world. We now had 15 Dublin Core metadata elements to work with rather than five ProCite web-software fields. The Tower project had never come close to using or needing all of ProCite's 40-odd fields, so we

were confident that Dublin Core's 15 fields would accommodate our data – indeed they seemed positively luxurious at that point.

The cooperative venture that followed included the following steps:

- We FTPed our four Tower databases (papers, memorabilia, photo/av and books) to Emory
- The MetaArchive staff mapped our ProCite data fields to Dublin Core and returned them to us for comment
- We conducted a clean-up of our database in order to make certain that 1) fields were used consistently both within each database and among the different databases 2) each item had a unique identifier 3) casual notes and comments meant for internal use were suppressed or edited
- We verified fields that were not used and agree to notify project staff if we needed to use any of these fields
- The MetaArchive staff "crosswalked" our data, which we then examined for completeness and accuracy

Throughout the process, there were numerous emails, conference calls, and telephone calls between our staff and the Emory staff. The process was not entirely painless since we were among the first institutions to submit data. We alternated between calling ourselves "guinea pigs" and "pioneers" as the MetaArchives technical staff and our staff traded professional vocabularies and knowledge. We learned about "crosswalks" and they learned about "series." The steps bulleted above did not happen with the snap of one's fingers, and this is no doubt truer for the Emory based staff than us. The FTP required some assistance from Southwestern technical staff on our end, and Emory eventually had to temporarily remove some firewalls to allow the data through. Data clean up was more tedious than difficult, but it had to be done meticulously. Southwestern University was very fortunate to have a graduate student intern, Holly King, from the University of Texas's Information School who handled this adroitly. ProCite's ability to search on specific fields and to make global changes also eased the undertaking.

Mapping the Dublin Core elements did seem quite straightforward from our end, and the 15 Dublin Core data elements appear to successfully accommodate records controlled at the box and folder level, or as is the case with the Tower Memorabilia database, at the item level. Although ProCite has over 30 templates (called "workforms") for every type of material from books to manuscripts to web pages and patents and these workforms contain over 40 possible fields, we found it expedient to design a 14 field custom "workform" suited to archival collections. These fields conveniently poured into the 15 Dublin Core elements—at least from my perspective. I did wonder if our use of ProCite's "index" field for series and subseries titles frustrated or confused the Dublin Core catalogers. Another concern I had was the lack of controlled vocabulary in our ProCite database. Dublin Core recommends controlled vocabulary for dates, places, subjects, media types, and so forth—and our use of controlled vocabularies was fairly casual. If, however, we had stopped to consult subject thesauri we would still be processing the collection. Meanwhile, our metadata is available now, and it is good and useful.

Sharing Metadata: The Success

Both the quality and quantity of information that this project makes available to researchers is stunning. For Southwestern University, the actual data crosswalk did happen with a finger snap, and I will leave it to the MetaArchives staff to detail any behind the scenes pain and suffering, if it existed. Meanwhile, researchers from anywhere in the world can now see detailed information about Senator Tower's papers at Southwestern University. Tower was chair of the Senate Armed Services Committee, and there is considerable interest in these materials, particularly in relationship to Vietnam. A search on "Tower" and "Vietnam" of the MetaArchvies database retrieves over 150 records from the Tower databases. I randomly selected two of the results of this search and printed them below to indicate the rich information now available to remote researchers.

- **Identifie**r 973-7
- **Title** Vietnam-1966
- **Author** Office of Senator John G. Tower
- **Subject** Defense/Foreign Relations/Armed Services Committee
- **Abstract** Pamphlet: "Viet Cong Use of Terror; A Study"; May 1966, U.S. Mission in Vietnam; Saigon, Vietnam
- **Discovery** Jan.-Sept. 1966

- **Identifier** 974-8
- **Title** Vietnam-1970
- **Author** Office of Senator John G. Tower
- **Subject** Defense/Foreign Relations/Armed Services Committee
- **Abstract** Bui Diem; reports: "Land Reform in Viet-Nam: A Historical Background," "President's Fact Finding Commission on S. Vietnam" (June 10), "Cambodia Concluded; Now It's Time to Negotiate" (Nixon,June 30); Nguyen Cao Ky; remarks, statements, press releases, fact sheets, Poll (RNC-Rogers Morton),etc.; draft (heavily edited): "Vietnam: Gradualism or Victory?"
- **Discovery** n.d, Feb.-Sept. 1970

In sum, the collaboration among the Emory based technical staff, the librarians who were Dublin Core experts, and our institution worked well for Southwestern University and enabled us to mount a more sophisticated and useful database than we could have done on our own. I anticipate creating an appropriate hotlink to the American South project from our online *Guide to the John G. Tower Papers,* and perhaps the American South can have a link to our website. While the Southwestern University based guide will provide background information for Tower's life and career, series descriptions, and an overview of the collections; the American South site will give researchers access to the papers at a detailed box and folder level. Between the two, one hopes the sum will be greater than the parts. I strongly believe researchers will think so.

The Future: Questions and Considerations

Archival collections are not necessarily static. There is every possibility that Senator Tower's family may give us new materials. Or they may subtract items since some are on loan. Even now, an indexing project is improving subject access to the Senator's correspondence. Will the American South project be able to update metadata to reflect changes in the data at partner institutions?

What is the long-term prospect of the project? Can partner institutions count on the American South project to be permanent, or might it evaporate along with funding?

Can the project coordinate with OCLC projects? Could a collection level record for the JohnTower Collection in WorldCat hotlink to my institution's data at American South?

Will partner institutions be able to add metadata for new collections as long as they relate to the currently defined areas of interest, such as the South, religion, politics, etc.?

In addition to importing existing metadata from partner institutions, each of which may use different software, would it be useful for the project to create a template that partner institutions could use for direct metadata entry?

Adding OAI Provider Capabilities to Digital Archives, Perspectives from the University of Tennessee at Knoxville

Anthony D. Smith and Chris Hodge, University of Tennessee at Knoxville

Summary

The University of Tennessee was able to develop OAI data provider capabilities by leveraging its strengths across campus. A partnership between UT Libraries and UT SunSite has allowed for the development of a data provider service to work as a harvesting point for OAI-PMH. This paper outlines the history of the UT Library/SunSite collaboration and the future trends of both organizations with regard to the newly-formed protocol.

Background

The implementation of OAI-PMH at the University of Tennessee is one of many recent collaborations between the UT's Libraries and its Office of Information Technology (OIT). While not unique, such collaborations are still notable in higher education. Collaborations are difficult, as Martin Halbert has pointed out, but they are not impossible. It's likely that UT would not have been able to participate in the AmericanSouth Project had it not been able to take advantage of the OIT/Library collaborative approach.

UT Libraries has been involved in digitizing its content since 1993 when it was awarded a Commission on Preservation and Access Grant to digitize the compositional works of Galston-Busoni. In 1999, the Library received grant funding from the Institute of Museums and Library Services (IMLS) to digitize Southeastern Native American Documents. In 2001, it received a second IMLS grant to digitize historical Tennessee documents from the state's antebellum period. Since 1999, the Library has been the recipient of over $1.3 million in external funding to digitize its collection and for the purpose of building a digital library infrastructure.

OIT's SunSITE program was created in 1995 through a grant from Sun Microsystems to encourage the adoption and implementation of emerging technologies and standards within the university, and to use technology as a means of fostering closer ties with the community and the private sector. SunSITE works within several broad areas where technologies are having an enormous impact on higher education: audio and video-over-IP; digital libraries, caching and replication; distributed education and lifelong learning; accessibility and the Digital Divide; the development and distribution of open source software; and the display of information in a geospatial and geo-temporal context.

When the AmericanSouth project organizers began formalizing their plans for content providers in 2001, they approached the UT Libraries. Despite its years of experience with digitizing, the library did not have the technical expertise or the infrastructure to establish an OAI-PMH service within the time constraints of the AmericanSouth Project. What it did have was a relationship with UT SunSITE and was able to approach the research and development unit and ask for assistance in building an OAI-PMH service that would provide access to the materials that had been digitized by the library. SunSITE staff was given the task of exploring the OAI-PMH standard and developing a data provider repository/service. The service would provide OAI records from the Library's digital collections as well as other UT-created materials relevant to the AmericanSouth Project. In order to accomplish this, it was first necessary to perform the translations for a variety of legacy metadata to the standard OAI-PMH schema that could then be interpreted by the harvesting agent. An open-source approach was employed to perform the task of establishing an OAI repository. Crosswalks (written in PERL) were created for TEI Lite, FGDC, and EAD formatted records to the necessary unqualified Dublin Core schema in XML format. The new records were stored using a MySQL relational database management system. MySQL offered a free open-source solution to processing and storing records. Three possible scenarios for storage were identified:

- Store each OAI record to a single MySQL data element field. This would offer quick response for a harvesting agent.

- Store each DC element in separate MySQL data element fields with multiple record elements separated by a tagging scheme. This format allows for a record to be entered once and retrieved in various formats upon request.

- Native XML file storage for archival purposes and to accommodate local search engine requirements.

Current and Future Developments: UT Libraries

After an initial test phase, the OAI-PH implementation was moved to the UT Libraries. This coincided with the Library's establishment of a Digital Library Center and its commitment to creating and providing OAI records for all digital library content. In 2002, the Library also purchased the University of Michigan's DLXS software to serve its digital library needs. In early 2003, the Library upgraded its DLXS implementation to version 10 in large part due to the built-in OAI broker service. The new service offers the Library an integrated approach to providing digital collection access at both the local and federated levels.

From the Libraries' perspective, there remain many questions with regard to the benefits of our investment in OAI-PMH. As the digital collections grow and OAI-PMH becomes a more mature service, it will be interesting to observe the content discovery process more closely. What proportion of users will approach the collection from an OAI-based cataloging service? The assessment and follow-up activity plan for the AmericanSouth Project will likely provide a wealth of insight with regard to the usability of the AmericanSouth federated access point. One would surmise that a gradual increase in the number of AmericanSouth users will occur over time as the service matures and is able to implement user assessment and as more become aware of the value of OAI/Dublin Core cataloged collections. The issue of using a controlled vocabulary and standardized rules for Dublin Core cataloging is a lingering question of whether the benefits of its use outweigh the cost of implementing. This development of a standards-based approach to multi-institutional metadata record creation is a difficult and complex hurdle to overcome

and may require time to evolve before OAI can effectively service the user in the discovery process.

In Tennessee, strategic planning for a state-wide digital library of cultural heritage materials has begun in earnest. The effort is being led by Tenn-Share, a non-profit organization that represents over three hundred library and information agencies throughout the state. In September 2003, the Preserve and Share Task Force presented its strategic plan and summarized the mission of the project as follows:

> Through the collaborative efforts of Tennessee archives, historical societies, libraries and museums, the residents of Tennessee will have online access to the visual and oral record of Tennessee's history, culture, government, and industry. The project will also assist institutions in the digitization process through training, tools, technical guidance and documentation.[1]

The plan calls for the University of Tennessee to serve as the host institution for the *Preserve and Share* cultural heritage collection. This strategy would rely on the OAI-PMH as a technical solution for building a statewide federated catalog of cultural heritage materials. The University of Tennessee would serve as the repository for metadata and possibly the content for some content providers. In addition, it would make the metadata available for harvesting through its OAI broker service. UT will serve as a harvesting agent for state agencies and may also simultaneously serve as a data provider agent for regional or national aggregators like it has done for the AmericanSouth Project. The University of Tennessee is poised to serve as a middle-tier OAI-PMH aggregator, which in itself, should provide some interesting observations about the exchange and sharing functionality of OAI metadata.

In the process of becoming an OAI-PMH data provider, The University of Tennessee has also been able to contribute its digital holdings for the purpose of federation in a number of different and useful ways. The Networked Digital Library for Thesis and Dissertations (NDLTD) is one such example of a federation effort that The University of Tennessee has been able to make a contribution too due to the development of OAI-PMH.

Current and Future Directions: OIT's SunSITE

Since the OAI-PH testbed ended, SunSITE's focus has shifted to exploring the implementation of metadata schemas for specialized datasets that at the same time are compatible with OAI. While these projects, at least in the short-term, are somewhat tangential to the direction AmericanSouth is taking, they may have a greater significance in the long term.

Through its participation in the Video Development Initiative's Video Access Working Group, SunSITE helped develop the metadata schema currently being implemented for the Moving Image Collections Portal (http://gondolin.rutgers.edu/MIC/), a project being developed by Rutgers University and the University of Washington on behalf of the Association for Moving Image Archivists. This schema permits the cataloging of materials in MARC21 and Dublin Core/MPEG7 formats. Features of the metadata records include:

[1] TEL II Preserve and Share Task Force, *TEL Phase II Strategic Plan: Preserve and Share Tennessee History and Culture,* August 2003, available from: http://diglib.lib.utk.edu/tennshare/tel2/docs/tel2draft_aug20.pdf

- Mapping to core data elements in a registry both to aid in record ingest and to provide consistent, interpretable search results
- Support for the archive's own data element labels and data element display order
- Extensible format-independent metadata design that accommodates searching, export and display in MARC21, Dublin Core, Dublin Core-Education and MPEG-7

SunSITE is currently exploring the feasibility of applying this same schema to instructional materials delivered through UT's Digital Media Service (another OIT/Library collaboration).

A second project with which SunSITE has been involved with for several years is the National Biological Information Infrastructure (NBII), a USGS-sponsored project, a broad, collaborative program to provide increased access to data and information on the nation's biological resources. This project is utilizing Content Standards for Digital Geospatial Metadata established by the Federal Geographic Data Committee (FGDC-STD-001-1998). A small subset of NBII FGDC metadata was exposed to OAI harvesting as part of the initial OAI-PMH testbed, and SunSITE is now exploring ways to make all metadata generated by NBII harvestable by OAI. We are also running a DiGIR server to participate in and develop biological portals similar to the Mammal Networked Information System (MaNIS) (http://elib.cs.berkeley.edu/manis/). DiGIR -- Distributed Generic Information Retrieval (http://sourceforge.net/projects/digir) – is a project to develop and test a protocol for single point access to distributed data sources, based on HTTP, XML, and UDDI. Finally, we would like to explore, through our association with the Electronic Cultural Atlas Initiative, the melding of cultural, biological and geospatial data.

The third initiative with which SunSITE has become involved since our work with OAI-PMH is DSpace. SunSITE, which has had a DSpace server operational since March 2003, is working with the University of Tennessee's Classics Department to develop a pilot project that will allow faculty to catalog both research and instructional materials directly through the use of a controlled vocabulary. Our primary interest in DSpace, which now supports only unqualified Dublin Core, is to make it usable with a variety of metadata schemas.

ARC – An Open Source Metadata Harvesting System

Kurt Maly*, Xiaoming Liu^, Moammad Zubair*

*Computer Science Department, Old Dominion University, Norfolk, VA

^Digital Library Research and Prototyping, Los Alamos National Laboratory, Los Alamos, NM

Summary

The Arc service is the first federated search service based on the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). The Arc open source system has contributed to the OAI-PMH protocol development and adoption by demonstrating the feasibility to harvest on a continuing basis from a large number of digital libraries. It has been used in a number of production services including MetaArchive.org, ncstrl.org, RDN, and snelonline. The Arc system represents a comprehensive solution for communities to harvest, index, and search, as well as for third-party service providers to harvest from Arc. The community involvement and open source process will continue to play a vital role in Arc's development.

Introduction

Many digital libraries have been built in isolation utilizing different technologies, protocols, and metadata in terms of both syntax and semantics. These differences hinder the development of digital library services that will enable users to discover information from multiple libraries through a single unified interface. The Open Archives Initiative Protocol for Metadata Harvesting (Lagoze & Van de Sompel, 2001) is one major effort to address technical interoperability among distributed archives.

Arc (http://arc.cs.odu.edu) service is the first federated search service based on the OAI-PMH. Arc was initially released as an experimental service to investigate issues in metadata harvesting in October, 2000. It has since then been used in a number of production and research projects.

The software developed for Arc service (http://oaiarc.sourceforge.net/) has been released as an open source system under NCSA-style license in September, 2002. It has been used in a number of production services including MetaArchive.org (Halbert, 2003), ncstrl.org, RDN, and snelonline. The Arc system represents a comprehensive solution for communities to harvest, index, and search, as well as for third-party service providers to harvest from Arc.

Development of Arc

Arc originates from the Universal Preprint Service (UPS) prototype (Van de Sompel, Krichel, Nelson, et al., 2000), which was developed as a proof-of-concept for various DL

technologies, including the feasibility of constructing a cross-archive searching service. UPS in turn is based on NCSTRL+ (Nelson, Maly, Shen & Zubair, 1998), a modified version of the Dienst protocol (Davis & Lagoze, 2000). Lately we re-implemented the core NCSTRL+ services using Java Servlets and an Oracle RDBMS. UPS became the foundation of the Santa Fe convention that in turn led to the Open Archives Initiative. Once the OAI metadata harvesting protocol stabilized, it was possible to realize the vision of UPS in Arc, with a higher performance search capability and its contents being kept up to date through the use of OAI -PMH based harvesting.

Arc was initially released as an experimental service to investigate issues in metadata harvesting. It immediately attracted interests since it was the only vehicle to demonstrate the potential and promise of OAI-PMH at that time. As new data providers appeared they often requested to be added in Arc for demonstration purpose; by continuously integrating various new data providers, the software was made stable and fault-tolerant. It has become a valuable tradition that Arc system tries to keep track as many as possible of the OAI data providers, so far Arc has harvested 6.4M metadata records from 165 data providers, since there is no centralized registration in OAI framework, this number is far from complete and we are continuously working on adding more data providers. Originally conceived more of a tour de force, Arc has become a useful tool for helping new data providers to make their collections truly OAI-PMH-compliant by giving them feed back on errors during harvesting. Secondly, it is becoming the 'Google' of the OAI world, however, at a cost: performance is becoming degraded as the collection approaches the 10 million record size while being maintained on a shoestring.

We are using Arc as the core engine for the two recently funded NSF projects: Archon and Kepler. Arc software is also used in the production NCSTRL (Networked Computer Science Technical Report Library). To apply Arc software and technology in various environments, many new features have been added for customization and installation of the system.

The Open Source Arc

In April, 2002 we were approached by MetaScholar.org for the software and license issues. Although we sent the software to several research groups before, the software is provided "as-is" and license has not become an issue. However for the MetaScholar.org project, an open source license was necessary.

After consulting with university administration the software was allowed to be released in an NCSA-style license. After cleaning up the code and documentation, the Arc system was released as an open source project hosted at SourceForge in September 2002. As suggested in GNU GPL FAQ (http://www.gnu.org/licenses/gpl-faq.html):

> A crucial aspect of free software is that users are free to cooperate. It is absolutely essential to permit users who wish to help each other to share their bug fixes and improvements with other users.

The MetaScholar project. and other interested parties have contributed bug-fixes, new modules, and documents. As a result of this feedback, the Arc software has become quite robust.

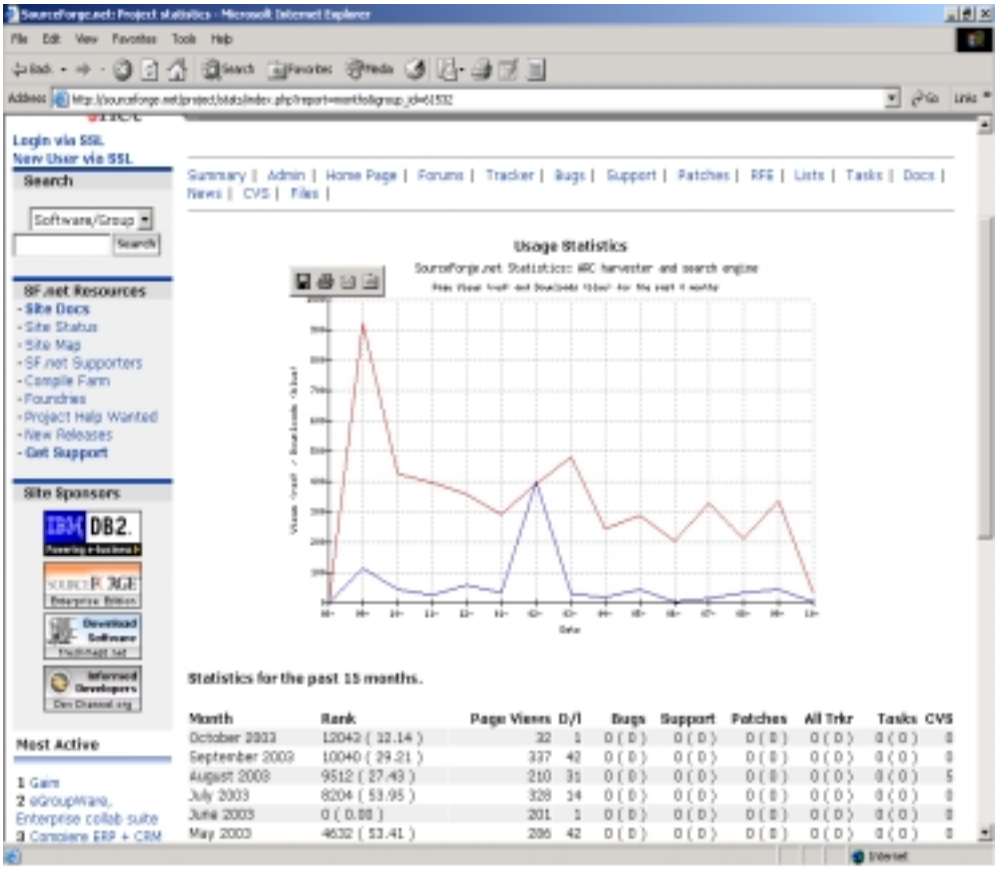


Figure 1. View/download statistics from Sourceforge

In Figure 1 we show the number of pages viewed and the number of downloads which remain fairly constant since its inception. One need to remember that Arc is not software for personal use but is established for a larger community that needs to federate several digital libraries. In that light the number of downloads is encouraging.

Features of Arc System

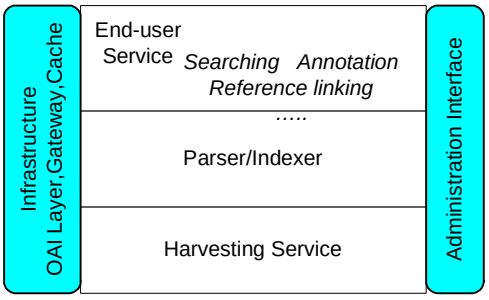


Figure 2. Architecture of Arc system

The Arc architecture is based on J2EE, moreover, the changes required to work with different databases are minimal. Our current implementation supports two relational databases, one in the commercial domain (Oracle), and the other in the public domain (MySQL). Figure-2 outlines the major components: Harvesting service; Parser/Indexer; End-user service; administration interface, and infrastructure.

Similar to a web crawler, the Arc harvester traverses the data providers automatically and extracts metadata. The significant differences include: normalizing the metadata to allow for complete and accurate searches and exploiting the incremental, selective harvesting defined by the OAI protocol.

The Arc parser/indexer service turns harvested data into internal representation for other services. The Arc parser is designed to be flexible to plug-in any metadata format parser. The Indexer uses full text indexing procedure of the underlying database. We are investigating other indexing/searching software as well.

OAI-PMH uses unqualified Dublin Core (DC) (Weibel, Kunze, Lagoze & Wulf, 1998) as the default metadata set, and most Arc end-user services are implemented on the data provided in the DC metadata. The current supported end-user services include simple search, advanced search, interactive search, annotation service, and browse/navigation over search result. The reference linking service as developed in the Archon project (http://archon.cs.odu.edu) will be integrated into Arc system soon.

Arc has a web based administration interface, which allows users to configure various parameters for harvesting and check harvester logs to handle various error situations such as erroneous XML replies from data providers.

Current Development Model

Currently ODU still plays crucial roles in the Arc system development. Various funded projects (archon, Kepler, TRI) are based on Arc system. New features are first developed and tested in these research projects, when appropriate they are integrated into Arc open source software and service.

To best demonstrate this approach we consider a case in the Kepler project (http://kepler.cs.odu.edu) that intends to build OAI-PMH compliant "archivelets" residing on a researcher's desktop. The archivelet is essentially unstable (in terms of availability to harvesting and searching) so an important features is to cache fulltext document in the service provider side. After the Kepler package is completely developed and tested, it will be integrated into the Arc system with functions to switch this feature on. Similarly many features from DP9 and the archon projects were also integrated into the Arc system.

The involvement from MetaScholar and other interested parties and individuals are playing vital roles in further development of Arc. The cooperation happens at various levels: bug report, bug-fix, suggestion, code and documents contribution. The vitality of an open source project is highly dependent on community involvement.

Deployment of an ARC System

The Arc system is based on Java/Servlet/JDBC technology. All required software is available at open source. Typical prerequisites are Apache Tomcat/Java/MySQL, or if the

user selects it Oracle. It is a pure Java-based system and has been tested in Window/Linux/Solaris platforms.

The software is designed to run a comprehensive solution, including harvester, indexer, search engine, and data provider. Using only selected individual modules it is possible but will require in-depth knowledge of the system. The software currently in SourceForge is well tested and stable.

We suggest the installation be done by users with experience of system management. The software package includes a README document. We also acknowledge Biju Vargheese for a nice installation document available at SourceForge. After successful installation, the software has all administration functions available as a web application.

A complete list of OAI-PMH related tools are available at http://www.openarchives.org/tools/tools.html, they provides solutions for many different situations and platforms. Arc provides an out-of-box aggregator and search solution.

References

Davis, J. R. & Lagoze, C. (2000). NCSTRL: design and deployment of a globally distributed digital library**. Journal of the American Society for Information Science,** 51(3), 273-280.

Halbert,M. (2003). The Metascholar Initiative: AmericanSouth.Org and MetaArchive.Org, **Library Hi Tech**, 21(2), 182-198

Lagoze, C. & Van de Sompel, H. (2001). The Open Archives Initiative: Building a low-barrier interoperability framework. **Proceedings of the First ACM/IEEE Joint Conference on Digital Libraries**, Roanoke, VA.

Nelson, M. L., Maly, K., Shen, S. N. T., & Zubair, M. (1998). NCSTRL+: adding multi-discipline and multi-genre support to the Dienst protocol using clusters and buckets. **Proceedings of the IEEE forum on research and technology advances in digital libraries** (pp. 128-136), Santa Barbara, CA.

Van de Sompel, H., Krichel, T., Nelson, M. L., Hochstenbach, P., Lyapunov, V. M., Maly, K., Zubair, M., Kholief, M., Liu, X. & O' Connell, H. (2000). The UPS Prototype: An Experimental End-user Service across E-print Archives. **D-Lib Magazine**, 6(2).

Weibel, S., Kunze, J., Lagoze, C. & Wolfe, M. (1998). Dublin Core metadata for resource discovery. Internet RFC-2413

Auburn Findings from the MetaScholar Projects: AmericanSouth and MetaArchive

Sherida Downer, Interim Dean, Auburn University Libraries

Summary

This paper will discuss three major areas of impact the MetaScholar Project has had on Auburn University Libraries. The first important aspect to be discussed is the responsibility and commitment the project requires of the library and the staff involved in the project. The issues of planning, financial and staffing considerations and ongoing support will be discussed from the administrative viewpoint. Secondly, involvement in this project has lead to local, regional and national collaborative projects. The collaborative opportunities the MetaScholar Initiative has offered libraries is important to note as a part of this paper, and finally, this paper will introduce the collaborative project Auburn University Libraries initiated as a part of the AmericanSouth.com Collection development.

Management Issues

Deans and directors of many research libraries are continually finding new opportunities for their libraries to become involved in national collaborative efforts. This can be viewed by the library staff as exciting and challenging, or sometimes overwhelming and frustrating. The importance of planning for a full-scale research and digitization project requires input and collaboration among your own library staff before committing them to a major project. Planning and budgeting for a major collaborative project are key elements to the success of the project. Deans and directors often see the opportunities for collaboration as an extension of the visions or goals they may have for their libraries, but the actual successful implementation of the project requires strong support and generous contributions of staff time as well and a supportive budget.

The first issue to evaluate is whether there is appropriate staffing. This is often hard to effectively evaluate, especially if you have very little personal knowledge of what will be required both in terms of technical knowledge and time commitment. Assigning the project to a person who has authority and is responsible for the duration of the project is probably the key to the success of the project. It takes a special set of management skills and technical knowledge to complete a project. Digital projects also have a tendency to required uneven amounts of time at different points during the project. Planning and technical setup may require tremendous amounts of time at one point, but after some of those initial setup requirements are accomplished, a different set of skills from a different group of people may be required. The experience that has been the most successful for our library has been to keep the communication lines open, be willing to listen, be supportive, and sometimes provide guidance or be willing to add additional staffing when necessary. This is often short-term staffing support that may come from others within the library or the technology center on campus.

Providing time and support for appropriate training is a key element for the success of the project. We have budgeted for training, including actual training costs, travel and time to learn. Once the technical staff and librarians involved realize you are willing to support them and their training needs, their self confidence and commitment to the work the project requires becomes more manageable.

Project Implications

Any major digital project will have ripple effects on the rest of the library, the systems support, and the campus. It is important to prepare and educate the other departments within the library about the project and give them regular updates on the project. Not only will there be a commitment of staffing and time, but it also may take additional resources to purchase the necessary hardware. That means there is less money available for other departments and they may have different views on the necessity of the project. It was crucial for us to make sure the cataloging department was well aware of the MetaScholar project because their technical skills were necessary to implement the metadata requirements. The fact that we were beginning such a large project using metadata meant that the catalogers in particular, needed to increase their skill level concerning different types of metadata. Of course the systems department was an integral part of the project, but not all of the staff in systems were working on the project. It was important to keep them all informed about what their colleagues were working on, especially when we found that we needed some special skill set one of them possessed who was not a part of the regular team working on the project. Keeping them informed provided much needed support and assistance on some technical issues and helped them become a part of the project. Of course, making sure the reference and information department was aware and supportive of the project was a necessity. We found that they could often provide information on which collections within our library might be areas we should consider for digitization. They are the librarians on the front line with the library patrons and they are aware of what kinds of information the patrons are asking for when they contact a reference librarian for assistance. Often their insight into the subjects we need to provide digital collections to support were the most fruitful and appreciated by the library users.

Developing the Plan

Developing the digital project plan is the key to the final success of the project and will require input and commitment from everyone involved. Developing a digital plan that reviews the collections, identifies what is to be digitized and how those items or collections support the definition of the project is the first step. Defining the metadata scheme to be used can become a real hurdle in the project. The technical expertise of the staff that will be contributing the materials must be considered and is as important as the contribution of the materials themselves. If the project will include accepting materials and metadata from small archives or museums, the level of expertise to support metadata input is often very different from that of a large research university library. This is also an area where new metadata schemes are constantly being developed. Planning for some technical response when the metadata format is changing is important. As was mentioned earlier, budgeting for enough support for personnel, time, training and travel is essential and must be acknowledged at the beginning of the commitment to the project. Any digitization project is piece driven. The development of a time line is important and there must be a measurement of output at appropriate times during the project. A monthly report including answers to questions such as where are we going, what is left to do and have we accomplished the stated goals within the appropriate timeframes should be addressed.

Opportunities

The opportunities the MetaScholar Initiative and the ensuing AmericanSouth Project have offered has led not only to successful collaborative projects, but has also afforded libraries the opportunity to share resources to a much greater advantage. The Open Archive Initiative (OAI) has been successfully implemented at most of the participating libraries and has added a whole new dimension to resource sharing. The excitement generated within the systems department and throughout the library when we were successfully OIA compliant was a real boost to the project. The more important result was that our valuable materials, which for years had been filed away and often unknown to the majority of the library users, could now be delivered in digital format to their desktop.

Successful projects such as these also afford libraries, museums and archives opportunities for fundraising, publicity and financial support. The collaborative efforts impress boards and presidents with the idea of cooperation and utilization of funds to the greatest advantage. It provides recognition at a national level and encourages further cooperation and the possibility for access to grant monies. But the most important outcome of the initiatives is the access to very special and often otherwise inaccessible collections of materials that can be shared on a national level. Libraries must remain active in these initiatives to remain viable in the information delivery world.

Transforming America Project

The opportunity for a regional technology collaborative project came as a result of the association the Auburn libraries had with the MetaScholar Initiative and the AmericanSouth.com Project. Auburn has developed and presently supports, a project entitled "Transforming America: Alabama's Role in the Civil Rights Movement." This is a project that has as its initial goal to collect the information on the civil rights movement that is available in libraries of all sizes, museums, archives and private collections in the state of Alabama. We provide a metadata template and server space and provide training when necessary. This site provides finding aids enabling library users and researchers to access information on the civil rights collections held in multiple places throughout the state of Alabama. The eventual goals are to include finding aids for civil rights collections throughout the Southeast, digitize many of them and eventually expand the subject to include civil war materials and re-title the project to be "Transforming America: From Civil War to Civil Rights."

Since we were interested in including all the collections regardless of the size of the collection or the institution, we realized we needed to apply the lowest common denominator for metadata. The project utilizes Dublin Core metadata and XML to create a subject based cross-institutional research portal. Because of our role in the MetaScholar Initiative we are also able to provide OAI compatibility. The elements comprise data useful to all possible partners and address the issues of collection building and partnership. As we developed the project we realized that many of the partners, especially those from museums and archives have a number of issues involved in their ability to participate. They were often limited by the cost of personnel & training; the loss of identity as the owner or collector of the materials; the possible loss of revenue since some of them depended upon access charges for survival; the lack of a web presence and even the possibility of access to a web presence.

The project is:

- Subject specific primary source research portal of Civil Rights collections in Alabama
- Virtual repository of distributed archival finding aids and museum collection guides
- http://www.transformingamerica.org

In order to support the contributors to the project, a template was developed which provides fairly simple design and labels to provide the necessary metadata for the collection. The link to the data collection form can be found at: http://diglib.auburn.edu/transam/form.html

Although this project is just getting off the ground, access and information regarding some very valuable materials are now available and harvestable within the guidelines of the AmericanSouth.com project, and for other digital collections around the world.

The Music of Social Change: Using OAI-PMH in the Creation of a Digital Memory Site

Katherine Skinner (Emory University)

Summary

Through the MetaScholar Initiative, we have so far experimented with ways to produce a model for the use of the OAI protocol in learning communities through collaborative efforts that involve library directors, curators, catalogers, scholars, and technologists.

Building on this model, the Music of Social Change (MOSC) project, funded by the Institute for Museum and Library Services, will investigate and identify ways to: 1) extend this collaborative model to involve museum curators and directors; 2) share resources with other institutions in order to make relevant and important information more visible to researchers; and 3) experiment with the applicability of the OAI protocol in creating a digital memory site on a focused subject area. This article describes five of the major goals of this MOSC research project.

Collaboration

The original projects of the MetaScholar Initiative, MetaArchive.Org and AmericanSouth.Org, relied heavily on the collaborative efforts of key groups: the creators of metadata (library archivists and cataloguers); the creators of metadata harvesting software (technologists); and the users of metadata (independent researchers and scholars). Each of these groups held particular responsibilities. The library archivists and cataloguers both helped to identify key collections within their libraries that would be of high scholarly value, and assisted the MetaScholar technologists in accessing the metadata associated with those collections. The technologists studied the various database and non-database systems used by the partner institutions, created cross-walks to transform their records into OAI-harvestable materials, and created a web-based portal that enabled users to engage with these materials and with each other in a learning community. Independent researchers and scholars worked to give the site direction, to provide contextual information and organizational structures for the metadata gathered through OAI harvesting, and to give the site a design that would make intuitive sense to other researchers.

In these initial projects, only one museum, the Atlanta History Center, was included as a partner institution. Through our work with the Atlanta History Center, we realized the importance of increasing communication between libraries, *museums*, and researchers in order to benefit all three parties. Museums are key repositories of research materials, but have significantly different operational priorities and practices than libraries. Museums infrequently participate in the development of union catalogs of holdings typical of library

consortia. Consequently, valuable museum holdings are often invisible to researchers who could benefit from using their collections. Likewise, libraries do not have the expertise held by museums in preparing exhibitions that provide broad contexts for topics—a skill that is extremely helpful in creating digital repositories of information.

As we enter the second phase of projects under the MetaScholar Initiative, including the MetaCombine and Music of Social Change projects (funded by the Andrew W. Mellon Foundation and the Institute for Museum and Library Services, respectively), we are studying the current division between museums and libraries, and are experimenting with ways to bridge the gulf between them. We are currently combining the efforts of museum directors, curators, catalogers, scholars, independent researchers, and technologists in order to produce a model for the use of the OAI protocol that can provide better services, involving research collections held by different types of institutions, to learning communities.

Only by keeping the lines of communication open and flowing between all of these parties can we realize the potential of the web as a learning tool. To this end, the MOSC proposal encompasses the collections held by fourteen museums, as well as library special collections. These libraries and museums will be able to collaborate in several new ways through the use of the OAI Protocol in this project. Museum databases will be able to interoperate with library catalogs by automated aggregation of records from disparate archives to create virtual online collections. Archivists, curators, and librarians will have freely available software to support the construction of subject-specific virtual collections. The project will provide the tools and training required by locally maintained databases and centrally maintained shared systems.

Sharing Resources

Smaller institutions—both libraries and museums in this case—do not often have the staff and the time to undertake comprehensive cataloguing. Likewise, they do not often have the resources to keep their library's technical systems up-to-date. In some cases, this results in lower-quality records or records produced in outdated formats—or, in the worst cases, no records at all—for the special collections holdings of small libraries, and the exhibitions and archival materials of small museum.

Large institutions can help to share the load of cataloguing for major collections at smaller institutions, particularly for those collections that are of clear scholarly value. The end—having the materials available to the people who need it—more than justifies the means—dedicating staff time to a project involving materials owned by another institution. Libraries often fall prey to the corporate mindset of competition, especially where special collections materials are concerned. This is detrimental to the mission held by most libraries, which is facilitating the dissemination of important information. The information, not who owns the collection, should be the central issue.

One of the primary goals of the MOSC project is to identify key un-catalogued collections concerning music and the Civil Rights Movement. Small museum and library archives will gain access to cooperative cataloguing services that can assist with the retrospective conversion of records that currently are held in non-machine readable formats. The OAI protocol will facilitate these cooperative cataloging services, allowing larger institutions to share staff time and expertise with smaller institutions.

Dissemination

A key goal of this project concerns an issue that affects archivists and researchers on a daily basis: dissemination of work through digital sites. During our work with the initial MetaScholar projects, we have seen that scholars justifiably remain concerned about their relationships with the Internet in two main areas: creating content and accessing content created by others. While dissemination has long been a goal of scholarship, researchers often fear that they may "lose" their claim to the materials they have produced if they post them on the web for other users. This can happen in several ways: if they post informational pieces, other internet users can "steal" their resources, plagiarizing them more easily through this format than through print resources. If they choose to publish a work digitally that they later want to publish in a physical text, their potential future publisher may not want that material due to its digital existence. As digital publications are still not admitted in most tenure-review processes, professors are particularly concerned with this later issue.

Ideally, the MOSC project will join the growing swell of digital initiatives that are creating web-based journals, research sites, and scholar-oriented discussion boards, and will serve as a legitimating source for web-based dissemination of scholarly materials. Some of the main purposes of scholarship, after all, are to distribute scholarly work to a broad audience and to facilitate others' use of one's work. This does not need to remain an age of jealously guarded research secrets due to the inevitable misuse of digital materials by plagiarizers. Rather, as with print materials, structures should be put into place (and enforced) that will guide users in proper and legal use and citation of these resources. With the proper controls in place, digital sites could allow researchers to share work-in-progress, making newly discovered resources available to other scholars, and encouraging peers to comment on the direction a new work is taking.

Contextualization

One of the major benefits the Internet can offer to such resources as metadata records from libraries and museums, as we have identified through the current MetaScholar projects, is its enormous capacity for adding context to research materials. We have experimented with this capacity in two main ways: 1) by drawing together related materials that have been physically stored in different research collections around the country into a new digital library collection, and 2) by inviting scholars to contribute short articles that provide a broad base of information about a particular record or group of records event or phenomenon.

In the Music of Social Change project, we will construct a new collection from geographically disparate resources, focusing on the music of the Civil Rights Movement. This subject area will not be limited to the holdings of any single institution or institutional form—during this first year, it will incorporate relevant collections from at least fourteen museums and a dozen libraries. In following years, it should continue to grow along with AmericanSouth.Org, which will "house" this collection.

Unlike many collections on music and the Civil Rights Movement, this digital memory site will not be limited to the song texts and music performed in direct protests, sit-ins, and rallies by social movement communities. We will also incorporate materials that are related to the songs—including biographies of the singers, information about important institutions (The Highlander Research Center, for example) that assisted in the dissemination of songs and techniques for their effective usage, and information about the events and

locations in which the singing took place. Related materials may also extend to information concerning the recording labels and recordings through which professional musicians participated the struggle (including jazz, R&B, and pop songs).

Following the model set forth through AmericanSouth.Org, professional researchers who focus on music and/or Civil Rights Movement materials, will produce short scholarly articles to provide a broader context for the items in this collection.

Activism

Perhaps my favorite reason for the MetaScholar Initiative in general and the Music of Social Change project in particular is that we are creating freely available software packages that will enable archivists, curators, and librarians at other institutions to support the construction of subject-specific virtual collections.

The need for these freely available (and easy to use) software packages cannot be stated emphatically enough. As we have seen in other digital initiatives, if we (the libraries) don't do it, someone else will—and they're likely to sell the result back to us, both in terms of content and in terms of the software that makes the content available. When this happens, we all suffer (except for the corporation who produces the software and content). Materials become more difficult to access, and without freely available software, developing new portals will pose major challenges for other researchers. For researchers who do not have official ties to a major research library that can pay subscription fees to corporate systems, accessing the portals created by corporations will not be a viable option.

Digital Media and New Forms of Scholarship

Lucinda MacKethan, North Carolina State University.

Summary

How can new forms of collaboration and scholarly expression, enabled by the web and scholarly portals such as AmericanSouth, help scholars doing research and publishing in southern studies? In the following pages I will attempt to answer this question through a consideration of the following:

1. The need for collaboration between archivists and scholars
    - in writing descriptions for digital content
    - in designing coherent finding aids
    - in considering digitalized images that present collection content
    - in confirming adequate technological standards
2. The advantages of publishing subject guides and reference materials online
    - Alternate ways of presenting information (experiments in organizing, the ability to reformat; Example: *The Companion to Southern Literature* state entries in relation to an online Southern literature subject guide and the guide in the *Encyclopedia of Southern Culture*)
    - Ability to revise, update, interject qualifying information
    - Inclusion of visual and auditory supplements. EX: Scribblingwomen.org
    - Ability to link primary and secondary materials
3. The value of digitalizing conference proceedings
    - Pre-publication as well as refereed papers on closely related topics dealing with cutting edge ideas and experiments with theory and practice (Example of Southern Writers Symposium)
    - Timely feedback and extended discussion of papers
4. The potential for breaking down interdisciplinary barriers at portals such as AmericanSouth

Introduction: Research and Publishing on the Web

Research. As a scholar specializing in southern literature, my research interests are interdisciplinary, they involve the use of a broad mix of primary and secondary materials, and they demand a constant updating of my knowledge both of archival and reference resources as well as recent literary and historical analysis, available in print journals, books, and digital media. My current project, a biography of a slave from Liberty County, Georgia, has involved locating wills, letters, Confederate cavalry regimental records, and census data from the 1820s through the 1870s; reading slave narratives and recent studies of this genre; grounding my writing in articles and books in such disparate areas as postcolonial theory, rice planting culture, and southern historiography; seeking reviews of an obscure novel published in 1899; and experiencing one of those wonderful serendipities: the discovery of a valuable handwritten memoir tucked away in a church museum. While this adventure has taken me down many a dirt road along the Georgia coast, and into small county courthouses as well as elegant mansions housing historical societies, the web has increasingly saved time, money, and general wear and tear – on me and my Camry.

Like most teaching scholars I know, I am now also much more committed to introducing students to the value or researching primary sources through archival web collections. I have found that my students can often learn more from primary material searches than from reading and quoting academic literary analyses. For instance, when I teach Faulkner's Civil War novel, *The Unvanquished,* I now send students to www.docsouth at UNC-Chapel Hill to find slave narratives, letters and diaries of plantation women, and documents related to the Civil War. The papers that they write involve comparisons between these sources and parallel situations in the novel.

For beginning students as for me, the links that the web makes possible between archival materials and the literature we study are pathways to new kinds of knowledge-making that connect past and present, fiction and history, imagination and fat-gathering, reason and memory. The chief challenge for me and other scholars as researchers is not that there is so much "out there," but that the materials are so scattered, often undifferentiated, too seldom annotated. As the technological possibilities for harvesting data advance, the need to develop better ways of displaying and evaluating resources becomes increasingly critical. The Metascholar Initiative can be a giant step forward in this area, as I will touch on in Point One below.

Publishing. My recent work as co-editor of the reference work *The Companion to Southern Literature* in addition to three decades of publishing articles and books and presenting papers in southern studies has brought me to a point where I can see how dramatically the web, and portals such as Americansouth, can change how scholars design, convey, and assess their work. Print publication will not likely be replaced by the opportunities for alternative formats provided by electronic publishing. Indeed sites such as Project Muse and The History Cooperative provide full electronic distribution of volumes of many of the best print journals in Humanities and Social Sciences fields, greatly enhancing the durability and accessibility of these outlets. The State online encyclopedia movement and Americansouth's forthcoming presentation of *The Encyclopedia of Southern Culture* are milestones in providing access to valuable reference materials through the web. Still, scholarly web portals can add other, much needed dimensions to the scholarly publication process, providing opportunities for scholars **1)** to experiment with less traditional organizational formats for reference and subject guide articles (Point two below); **2)** to encourage new scholars through a pre-publication and review process; **3)** to capitalize on the focused sharing of ideas that takes place at scholarly conferences (Point

three below). Finally, scholarly portals can create a gathering space that frees knowledge seekers at all levels from many kinds of artificial barriers that abound in academe: those between early learners and professionals, those between scholars and archivists, and those between scholars of different disciplines, all of whom have common interests that should be nurtured through venues that promote the causes that they share (Point four below.)

1. The need for collaboration between archivists and scholars

Because my own experience in working directly with archivists in the areas listed below is minimal, to say the least, the most that I can do in this section is to express support for possibilities for collaboration that are certainly in place already in some institutional environments. My questions are guided in part by the goals of the MetaScholar Initiative, which provide a natural opportunity for interactive discussions about the nature and functions of Metadata. Lorraine Normore, in a web article entitled "Studying Special Collections and the Web: an Analysis of Practice," identifies the critical importance of such discussions (See http://www/firstmonday.dk/issues/issue8_10/normore/index.html). Archival collections, she says, are

> not just a set of items that have some commonality that binds them together. [They are] our cultural heritage, the story of our past, reflected in the things that were made by natural or social forces. If that past is to be kept, metadata about the entities need to keep both information about the thing and about the physical and cultural context that it is derived from. Metadata provides both needed information and a way to support access to the material and to the framework of knowledge that it resides in.

The imperative that Normore defines for metadata design is one that concerns archivists and scholars as one community, as do a host of other related questions: Once digitally encoded items have been harvested or made available for harvesting, how can and should they be presented? What minimal standards need to be in place when considering the acceptance of digital materials, in terms of both technology and in terms of the format and quality of information in catalogue records. What kind of descriptive commentary should accompany digital materials, can or should such materials be enhanced, and if so, how? What kind of finding aids are available and how accessible are they? How important are digitalized images depicting collection content? In what situations are web exhibits, such as the extensive one at the Valley of the Shadow site, appropriate?

At the Americansouth site, the archival collections of more than three dozen libraries, colleges, and museums will exist side by side with ongoing interpretive scholarly work in related subject domains. Yet existence side by side is not necessarily existence together. Will it be possible for scholars to generate interpretive assessments of specific bodies of archival materials in response to priorities identified by archivists? Will it be possible for archivists to consider creative kinds of groupings and displays of digitalized collections in response to agendas identified by scholars? Can differences between these two agencies, much less differences between institutions, academic disciplines, donors to collections and recipients of them, be bridged in the matters of standards, tools, accessibility, and interpretive frameworks? The stakes, as I think Lorraine Normore's assessment makes clear, are high, and the potential is certainly strong for full discussion of these questions at AmericanSouth.

2. The advantages of publishing subject guides and reference materials online

- **Alternate ways of presenting information (experiments in organizing, the ability to reformat)**

  AmericanSouth hopes to publish definitive subject guides to survey several areas of scholarship within southern studies, including the literature of the South. Such guides abound in various print resources, from book-length works such as The History of Southern Literature to lengthy entries in reference works such as The Encyclopedia of Southern Culture. My charge in thinking about how to design an overview of southern literature for Americansouth was to come up with one that would not only identify key texts but also point to a fruitful research emphases. The advantages of a non-chronological narrative emphasizing genres in relation to the concept of Regionality seemed to me to be an exciting alternative to the usual format. In this way it would be possible to explore the formal relations between literary works by grouping them according to a set of what I would call regionalized generic conventions – which are always socially and politically mediated For instance: southwestern humor tales from the antebellum period and what we call "grit lit" novels of the late 20th century are miles apart in chronologically ordered overviews, but in terms of language, character-types, and ideological agendas, they relate quite closely, and an overview that creates a Generic heading inclusive of both of these forms can suggest new ways to read and study these works. A site such as Americansouth can experiment with this kind of grouping without being locked into it and can even set up ways to switch to other models of overviews. We can send readers to an author or chronological overview such as the one provided by the Encyclopedia of Southern Culture, or to a state-by-state overview such as one available in The Companion to Southern Literature, which might be brought into the site. Or we can re-format the Genre Overview at one click to spread out headings in a different design.

- **The ability to revise, update, interject qualifying information**

  The publication of a reference work such as *The Companion to Southern Literature* was a mammoth undertaking, and it was very satisfying to see the one thousand pages and 500 entries by 276 different authors set in beautiful print between the hard covers of a book. But even before the type was set, it was clear how valuable some process of updating and revision would be. Indeed, I asked the publishers early on to consider publishing the book in some kind of looseleaf 3-ring binder format so that we could create supplements, appendices, different kinds of commentaries, etc. The reaction was predictable. In this respect, the beauty of web reference sources is clear. Since 2001 there are new texts that exemplify the themes, motifs, and events that are headlined in *The Companion*, to say nothing of new subject entries that should be considered and research that would change how some subject headings are defined. It is frustrating even now to admit that we didn't ever find a completely coherent rationale for what we included and excluded in our book – we simply ran out of time and space and money. Thus we have entries on horses and horse-racing, on the mule, and yes, on the pig in southern literature – but to the horror of canine lovers everywhere, no entry on the dog. Of course the flexibility that allows both reformatting, revision, and supplementing means a loss of stability in some respects. Reference works need some time to establish a viable and stable point of view before they come under the wrecking ball. But the ability to add, subtract, review, and respond reflect what research and reading are all about: the constant interchange and advancement of knowledge that benefits from assertions of dominating frameworks but also from regular reexamination and readjustment. With quality control through editorial boards and focus groups, it should be possible to encourage both flexibility and the promotion of "best views" and "best practices." It is safe to say that our expectations of the shelf life of dominant paradigms for interpreting and organizing knowledge are being adjusted

downward dramatically, and we can need to be engaged in multiple processes for assuring quality while not short-circuiting the processes of competition and paradigm shift as we make decisions about scholarship and review.

- **The inclusion of visual and auditory supplements**

  One of the great benefits of being on the Scholarly Design Team of Americansouth is the expansion of my horizons that has resulted from being connected with four scholars whose work is so relevant to mine but who come to the South and its riches from such different routes. The work of Allen Tullos should be mentioned in particular here because his encouragement of visual and audio dimensions to our presentations so aptly and dramatically captures what the web is all about. Certainly for his own area of music, the ability to reference the sounds and sights of the South is not just an accessory but a necessity. Yet for literature as well, these dimensions are points of access that should not be ignored.

  Digitalized voice and sound. In another project that I have been directing for several years, www.scribblingwomen.org, we present half-hour radio play adaptations of American women's short stories along with background, interpretation, biographies, and lesson plans. The radio play is both an old and a new genre with great potential for teachers and students alike, and to hear the plays in connection with several kinds of contextualizing commentary provides a rich experience. Different kinds of audio dimensions can enrich Americansouth. For instance, we have the opportunity to present texts transcripts of interviews with important scholars and writers, accompanied by a digitalized portion of key sections of audiotapes. Given that oral literatures and the use of dialect in print literatures are such an essential element of the South's literary heritage, the ability to offer examples – of folktales, slave songs and spirituals, and regional dialects from Ocracoke to Appalachia – can make a huge difference in how researchers approach literature. Visitors to Americansouth can match my discussion of Appalachian writers such as Lee Smith and Mary Murfree to Allen Tullos's display of southern Appalachian ballads and a Doc Watson jam session and hear for themselves a demonstration of the Celtic roots of both the literature and the music of this subregion. **Digitalized images.** The opportunity to balance textual content with images, including images that provide the original state of certain items such as wills, letters, and manuscripts, is especially meaningful to my work on Liberty County, Georgia. It is one thing to read a fictional account based on the actual division of the estate of Moses Liberty Jones among his eight children. It would be another thing to have a digitalized transcript of this division showing each child's portion represented with a "Lot" number followed by forms of property, including slaves, that were the child's inheritance. It would be quite another experience – having something of the compelling force of eyewitnessing -- to see on a screen an image of the listing in its original form: the handwriting, the matching of the slaves' ages, occupations, and dollar value, the rows of figures and items painstakingly recorded.

- **Ability to link primary and secondary materials**

  One of the most instructive and valuable outcomes of my work to produce an overview of southern literature has been my effort – urged by my colleagues on the SDT, to match my discussions of exemplary texts to E-Texts available through sites such as UNC-Chapel Hill's doc.south and the University of Virginia's Xroads where, for instance, many hard-to-find nineteenth century novels, short stories, and autobiographies have been assembled. Hypertext links are invaluable. Sending reader's to "the horse's mouth" so to speak, combining interpretation with illustration, provides a kind of balance that truly makes the Web into a dynamic classroom, to say nothing of saving hours of searching and waiting for interlibrary loan. A commitment to matching secondary readings, guides, and commentary

with relevant primary sources in archival collections can be a major benefit of portals such as Americansouth that will have such a rich balance of these materials along with the guidance of archivists and scholars to provide roadways and carefully considered selections.

3. The value of digitalizing conference proceedings

Most of my own publications began as conference papers, which up to now have been a primary method for scholars to try out and validate ideas, receive helpful criticism, and balance their own insights against others in an open environment of exchange. The pathway from conference presentation to journal publication to book has been natural and fairly inevitable. This system provides a good range of benefits: public recognition, dependable, long-term access, extensive editorial procedures, validation by peers. Yet managing scholarly information with the help of digital technology is a major imperative, given the challenges scholars, librarians, and publishers face today.

The editors of the eScholarship project at U California, in describing their eScholarship Repository, first list some "imminent threats to the sustainability of scholarly communication": (See http://escholarship.cdlib.org/about.html):

- the increasing volume and escalating costs of traditional scholarly communication
- economic and organizational barriers to entry in digital publishing or the creation of innovative alternatives
- adequate signals of quality within burgeoning literatures
- efficient dissemination among peers, with opportunities for review and commentary
- protection of intellectual property
- enduring availability for the future

One way to meet the challenges described in this list is the replication at scholarly portals of some of the best features of the scholarly conference model, including: well-defined topics that provide coherence to discussions; clear standards for quality in content and style maintained by Conference boards and session leaders; grouping of papers in ways that insure multiple perspectives and an atmosphere of dialogue; opportunities for questions and comments by audience/readers.

At the Americansouth portal it will be possible to produce online conferences but also to take advantage of actual conference proceedings. This September, the Southern Writers Symposium held at Methodist College in Fayetteville, NC, and now in its 17th year, chose the Concept of Region for its overall theme. Through an arrangement with the conference director, the papers already carefully selected to be delivered at this conference will go through an additional review process provided by two volunteer readers, keynote speakers John Shelton Reed and Jon Smith, and through an additional editing process provided by the conference director and me. The finalists' papers will be posted at Americansouth with opportunities for responses. The papers represent a wonderful range and variety of creative approaches to the concept of Region, from "Magical Realism and the Mississippi Delta," a paper that explores connections between Latin American writing

and the South, to “Swamp Tours and the Commodification of Southern Landscape,” and “Aristocrats in Blue Jeans: Appalachian Writers Since 1950.”

In taking advantage of its capacity to offer eScholarship opportunities, portals such as Americansouth provide rapid, timely access to new research, quality and editorial controls, helpful groupings of materials, and the benefits of peer critiquing and promotion of dialogue.

4. The potential for breaking down interdisciplinary barriers at portals such as AmericanSouth

Many prestigious print journals are often controlled by their disciplinary sponsors, and while the lines that divide history from literature, music from religion, African American studies from traditional Southern Studies, have been challenged and blurred, opportunities to bring multidisciplinary perspectives to bear on questions of scholarly importance are still all too infrequent. Even less frequent are concerted efforts to forge partnerships of librarians, scholars, and information technologists, and last but hardly least, their institutional funding administrators. Americansouth, having put this rare model of partnership together, is in a prime position to create an innovative and interactive online community whose interest in the culture and history of the American South is both expansive and inclusive. Such a community will be capable of keeping up with both the shifting boundaries of the South itself and the shifting boundaries of all the disciplines that comprise the study of how the region came to be and where it might be headed next.

In the Halo of the Moon: Significance of AmericanSouth.Org for Research

Erich Kesse (Digital Library Center, University of Florida)

Summary

These comments consider research as a dialogue between resources and their users. OAI is viewed as the first of a series of toward a desired research utility. Next steps and fantastic possibilities are considered, among them enlarging the field of contributors, targeting qualified metadata and additional data types, and enriching the interfaces for research utility.

Introductory Remarks

I am not a southerner. I should offer this much in confession and preface to these remarks on the research significance of AmericanSouth.Org.

I am not a southerner, though I can trace my *ruts* back to the foothills of Kentucky. My maternal grandmother used to say of an impending storm, " *'Pears a holler up the road, it's a-fixin' ta come up a cloud!*" That much I still readily understand. But, the meaning of some words is nearly lost. "*Seems we yar stars*," she intoned with sounds my vocal chords cannot find. "*Seems we yar stars wethin a hayla of the moon*." Perhaps, it was the poetry of a people, too long isolated. "*Wrench a cup, boy*," she demanded, "*an fetch me some rainwader when it comes a cow peein' on the flatrock of the rough*." I can recall the first time I heard those words. Cows jumping over the moon was the stuff of children's stories, sure enough; her way of speaking to a five year old boy. But, did she intend to say that the storm would be so bad that it would heave a frightened cow onto her roof? Perhaps for lack of understanding, my father sent me to live on a dairy farm in southern Indiana the next summer. There, I came to appreciate the manners of cows and comparisons to a hard rain. And, there, I lived among other migrants, like my grandmother, who had crossed the *Ohiya* but continued to count the days to a good soaking by the number of stars they could count in the halo of the moon.

It seems a fitting analogy, in looking to the significance of AmericanSouth.Org today, to say that we stand in the halo of light that both embraces and illuminates a distant but familiar object: southern cultural heritage. Today, we've spent a considerable amount of time talking about the moon. In bringing together this cultural information, it seems fitting also that we should take stock of why we shine a light on the moon.

Sin-eater: The Process of Information

In plain language, today, we have looked at OAI and AmericanSouth.Org, in particular, as an information gathering and warehousing technology. We have talked about it in the way that farm equipment engineers talk about the latest in combine design. There's nothing wrong with that. Efficient harvesting of information saves the researcher's time. And, isn't that a fundamental law of library and information science? Today's discussions have been a kind of Sunday among the Shakers.

But, now, we need to consider selling the harvest! To be certain, there is a fair amount of marketing to be undertaken. With the launch of new interfaces, the integration of an *Encyclopedia of Southern Culture*, and the commissioning of articles, now is the perfect time if not to sell then to consider what folks will do with the harvest. Of course, they'll consume it. But, what of that? The fundamental question is what do we do with the energy we gain of eating? Or, plainly, what will people do with this information harvest?

Two of the folk customs that most fascinated me among my grandmother's Pandora's-box of southern wisdom were the bottle-tree and the sin-eater. (I expect these will have prominent place in the *Encyclopedia*.) The bottle-tree, literally, a tree hung with empty bottles, bestowed voice upon the spirits,[1] their sounds whistling across the mouths of bottles. Every OAI project – AmericanSouth.Org, RLG's *Cultural Materials*[2], the University of Illinois at Urbana-Champaign's *Digital Gateway to Cultural Heritage Materials*[3], the University of Maryland and the Internet Archive's *International Children's Digital Library*[4], and others, is a tree hung with bottles.

Ignoring the more disturbing purpose of the bottle-tree[1] for the moment, I am proud to say that the University of Florida (UF) has its bottles hanging in several trees. OAI gives us an uncanny ability to be heard well beyond our primary market, the UF campus, where our resources, our produce are consumed. This most obvious significance of OAI is appreciated by faculty who understand that our contributions encourage others to bring forth their goods. OAI is a grocery store with resources brought to them by whatever means; they don't care. Its importance rests in the proximity, volume, and diversity of the collected resources. These information objects, from hither and yon, are literally at their fingertips. – But, this is to say nothing of the information process.

Love is a sickness, boy, my grandmother would say. Others have called it *hunger*. Both *sickness* and *hunger* are words ascribed to the sin-eater, a kind of middle-earth Santa Claus. A sin-eater eats anything you leave him, whatever the symbolic sickness or sin you've baked inside. Sin-eating is one of those professions that have a most immediate vitality. The engine of language hasn't removed it from the verb, from the thing that it does. The objects and the process are conjoined. A sin-eater *eats*. It seems to me that AmericanSouth.Org, less so than other harvesters, lacks a particular vitality. Harvesters don't have sin-eaters, for a start. They harvest, certainly. But, they've got information scientists and librarians. And, we describe and order other people's information. Harvesters need someone to make something of the produce filling the stock rooms they lay open.

The prospect of AmericanSouth.Org, more than any other OAI harvester, except *maybe* the International Children's Digital Library[5], is that it might teach directly through its articles, *Encyclopedia* entries, and other contributions generated through use of the harvested resources. So, in the life cycle of research, the process of information, AmericanSouth.Org has brought us to the harvest and has begun baking the possum into the pie.

Flatrock: the Monolith of Southern Culture: A Tool for Understanding and Collection Development

AmericanSouth.Org lends the impression of circumscribing a monolithic culture. Yet, we need only review linguistic and dialect diversity maps of American English[6] as spoken in the South to remind ourselves that the monolith is not homogenous – or *sweet* as my grandmother would have it, *sweet as possum pie*[7]. AmericanSouth.Org's articles, in particular, reflect the region's cultural diversity. But, let's, for the moment, assume that southern culture is a flatrock, a monolith that is somewhat homogenous.

Bittersweet, the demographics of the American South are such that southerners already have been out numbered within this fast growing region. The southern culture of American history that ends with my grandmother's generation might be made to seem a closed chapter. Digitization of southern cultural materials, too, sometimes seems akin to restoring a Mayan temple. We build on the evidence of a found architecture, allowing each new block, each digitized object, to find its place, sometimes by suggestion. Many of the several southern statewide digitization programs have constructed admirable temples of culture with the guidance of humanists, scientists and educators. No doubt, some of the pieces are out of context but none are without purpose. We hope to preserve the evidence of our past and to facilitate its use for the benefit of the present. All teach as well as gather resources; some better than others. Most of the learning modules[8] are a monologue rather than truly didactic. The common sense of country folk suggests that one sows seed following a harvest. Learning modules, articles and *Encyclopedia* entries should engage the researcher in the (harvested) collection, an engagement that should result in the creation of new objects.

To the extent that AmericanSouth.Org begins to outline southern culture, we builders of state and local collections can look to it to more easily see the lacunae among our own collections, constructing both better local collections and, in the process, a better, more complete regional collection. The topical approach to collections, available in former versions of the AmericanSouth.Org interface, will presumably be restored with the development of the *Encyclopedia of Southern Culture*. The former version listed resources under gross subject headings. With topics the likes of "Agriculture", "Geography", "Literature", etc., they were too broad for quick analysis by either collection builders or patrons. Assessment of the coverage for a particular religious movement, for example, was difficult. The *Encyclopedia*, I would hope, should provide contextual narratives and refine topics to lower levels of aggregation. Each should be a little nut of learning to explain the resources gathered under it. And, each should offer value to young researchers at levels K-12, if not to researchers at higher levels or to collection builders themselves. Topical analysis and continued development of topical collections will then be guided by knowledgeable interpretation from the regional perspective.

Missing, yet, are the tools needed to statistically analyze topical strengths and weakness, one digital library provider to the next. Everyone knows that a measuring cup is requisite to baking the perfect pie.

For the benefit of its users, I trust that *Encyclopedia* entries will be linked with the research resources located by AmericanSouth.Org for the topic. Currently, a glance at the articles commissioned by AmericanSouth.Org quickly reveals links to electronic resources – web pages, mostly – beyond AmericanSouth.Org. Articles infrequently reference the harvested collections. It is a curious indicator of AmericanSouth.Org's success that unlike the majority of harvester websites it provides wholly new information but does not reference the information it, itself, has collected. Curious, too, that it comes at a time when

K-12 educators struggle to teach their charges evaluative techniques[9] that question the value of the many unvetted Internet resources. Internet resources cited by the articles, by the way, are of excellent caliber. Perhaps, we assume too much of our audience.

Monet rather than Monolith: Inspiring Research

Linking articles and *Encyclopedia* entries to the content of AmericanSouth.Org suggests opportunity for the enhancement of OAI. Automation of this task is an area for development. Subject assignment is an art practiced differently, with varied thesauri and descriptive practices, among the collections harvested. Reports from nearly every harvesting project, AmericanSouth.Org among them,[10] have described metadata as heterogeneous. Some librarians have privately suggested that, regardless the care taken in subject assignment, patrons just don't use the same language in searching library collections – digital or otherwise. And, at least one OCR engine[11] even hides variant word forms and common mis-spellings within the bitmap references/tags for individual words – all for the benefit of discovery by the widest possible audience. While this takes us far a-field, to another harvest, it seems to suggest that what we strive for with OAI stands at the gates of research support.

Topical searches across AmericanSouth.Org's harvested resources are limited now to keyword searches of the too-limited information gathered. Perhaps, scholars authoring *Encyclopedia* entries might eventually find themselves working with programmers to develop analytical tools that process harvested records to automate topic assignment. Or, perhaps less ambitiously but more immediately, they will work together after the harvest to normalize resource descriptors – the subject headings – used in the harvested records. I envision, in effect, the creation of an automated means of sorting the various harvested resource records into virtual peas and corn, or avocados from oranges, a means of associating resources with *Encyclopedia* entries.

For the researcher who relies upon point-and-shoot methods, with or without the cross hairs of a thesaurus, more information increases the odds on successful discovery, engenders more comprehensive research and, ultimately, makes possible more thoughtful teaching. AmericanSouth.Org, like the digital libraries upon which it relies, is more a Monet than a monolith. The closer you get, the more apparent that nothing but mission and programming connect the dots. The observation looks both ways. AmericanSouth.Org needs to increase the harvest yield both in terms of contributions and the amount of information collected. More providers -- more information -- the more able, the researcher to shed new light, to find new interpretations, to tell new stories that teach and enlighten.

SOLINET's Kate Nevins, during a June 2003 project-planning meeting, suggested that AmericanSouth.Org wants for expansion, to encompass college and public libraries, museums and other cultural institutions across the American South if it is to more nearly characterize southern culture. Research, after all, is in large part a search for information. And, the smaller institutions often hold, in relative obscurity, pieces of a puzzle. Consider research in to the life and eventual insanity of Mary Todd Lincoln, First Lady, Kentucky belle. AmericanSouth.Org harvests little information on Mrs. Lincoln. We find Elizabeth Keckley's scandalous tell-all *Behind the Scenes*, which was removed from sale shortly after it was published in 1868. But, the researcher is left to discover elsewhere and independently letters of a younger Mrs. Lincoln. One letter written, seemingly in blood drawn by a sharp tongue, describes the curiously insane burial of a honey-glazed ham.

The letter, hidden among the archival collections of Lexington, Kentucky's Transylvania University, *alma mater* of Todd men, underpins the rage of a woman all but abandoned.

Similar jewels, some grand, others small, are found in smaller institutions across the South. At Centre College in Danville, Kentucky, Doris Betts, a North Carolina writer, can still be heard advising writers of southern fiction; but, no reference to her can be found in AmericanSouth.Org. And, AmericanSouth.Org is blind to the complete and completely digitized WPA *Florida Writers Project* collection at Jacksonville University in Florida. For the researcher not familiar with AmericanSouth.Org's fellowship, AmericanSouth.Org is one church among a multitude.

For those of us now contributing, there is more, as well, to be given to southern research studies. The University of Florida, for example, is only now acting to open the Spanish borderlands collections. My state's history cannot be told without Spanish and French voices or British accents. The first Confederate state subdued, the only Confederate government not to have fallen to Yankees. Florida was a contradiction well before the last presidential election. The import of Spanish and British opinion on slavery through the acquisition of the Florida territory was reviled throughout the American South prior to the War Between the States. And, yet, these opinions are largely unknown to AmericanSouth.Org.

Providing more information is requisite to the research success of AmericanSouth.Org, but doing so means instituting the Project more widely. As Emory University acts to extend the utility of the harvester in processing information, perhaps it is incumbent upon institutions to seek partnerships and funds that will bring their resources into AmericanSouth.Org. A challenge for us here, today, and for scholars seeking funds is to share our knowledge about repositories though OAI for the benefit of future research.

Stones on a Harvest Moon: Qualified Metadata

I see the greater research-supporting role for Emory and its technology partners to be in extending OAI to harvest more qualitative or, rather, qualified metadata. In the field of research, rich metadata yields a stronger research return. OAI needs to dig deep to more aptly support research.

Among the PALMM Collections[12] of Florida's cooperative digital library program, the University of Florida and our partners have elected to build under shared technology with common standards. We ingest rather than harvest. This strategy provides us with access not only to the basic bibliographic metadata commonly harvested under OAI but also to richly qualified metadata, using a variety of multi-layered thesauri, codes and classifications, and other knowledge systems. And, with implementation of new, fast and highly accurate OCR in combination with semi-automated tagging scripts, the strategy will soon allow our researchers to dive into a larger number document texts as well.

Until OAI, targeted query engines, and other systems are adept at, first, query of qualified and structural metadata and, second, query inside of tagged finding aids and text, I would hope that PALMM would continue to ingest. The researcher is limited within harvested collections by the small set of information *now* gathered. These limitations suggest future venues for development by Emory University, the Universities of Illinois and Maryland, the Research Libraries Group, the National Science Foundation, and other harvesting agencies. Development of harvesting for qualified metadata and tagged text, of course,

would also include extending the benefits of the additional harvest yield to research in the form of new utilities.

A harvest of qualified metadata will certainly reap everything that the Dublin Core Metadata Initiative places under its stylized logo sun. Two of these added elements seem to me to be more essential than others on the periodic table of research: time and space. No metaphysic of information can be complete with out them. History is a series of events spread over time and space. And, in no American region is time and space more important than it is in the American South.

> When I moved to Florida, an elderly gentleman would grab me on my nightly walks around the neighborhood that once had been the University of Florida campus. He would hold me by the elbow and guide me from block to block as though I might have been naughty child. In front of each house, he'd shuffle to a stop. *Miss Emma lived here when I was a boy*, he might say. *She was an upright woman. 'would boll-up peas she planted herself. Don't ya know, her husband, Jimmie, was a no account …"* And of course, I didn't know. The marvel of those walks, night after night traveling the same blocks, was that the stories were never the same. Each night, he would bring the neighborhood's characters back to life in a sequence spread like peanut butter over time.

Doris Betts, the North Carolina novelist, once remarked[13] that though she researches the motivations of her characters, be they doctors, lawyers or common folk, moving them across a room in time to discover the details of plot is the most challenging aspect of work. Plotting data on time-layered maps – whether tangible or held in the mind of the researcher in the course of work -- has been the labor of linguists, historians, scientist, and writers. Eventually, it should be ours as well if we intend to serve their research interests.

The anticipated utilities go beyond OAI in that they will not be about simply harvesting information. They will be about manner of using harvested data. Several projects point the way. The Alexandria Digital Library's Gazetteer[14] at the University of California – Santa Barbara (UC-SB), for example, utilizes a harvest of geographic data pointing to maps. The Perseus Digital Library at Tufts University[15] plots the location of resource repositories, using the language of Space/Time. And, the recently IMLS-funded *Ephemeral Cities* project among the PALMM partners in Florida, promises to pilot a harvest of Florida's cultural information and artifacts as historic layers into a geographic information system (GIS). We trust that the *Geo-Temporal Core*, as we call it at the University of Florida, will be a reactor and a catalyst for research and, particularly, for developmental histories. The concept is simple in theory: isolate the geographic and temporal metadata in our records and tagged text, normalize it, and subject it to query by text, by timeline, or by map interfaces.

Nuck. Nuck Muck!
Beyond the Language of Old School

Searchable text documents remain a challenge for harvesters as well. Access to machine searchable text is a holy grail among researchers seeking to save themselves time.

To some extent, OAI is antiquated. It does for digital libraries what the *National Union Catalog* and the *National Union Catalog of Manuscript Collections* did for brick-&-mortar libraries. At the moment, OAI is a divining rod for researchers. It now needs to dig deep, to be a well. – I say this with the full understanding that wells are dug *one shovel at a time* and that Emory and other harvesting agents are now struggling with qualified metadata.

But, as a first step beyond qualified metadata harvesting, OAI might harvest structural metadata – tables of contents – perhaps using METS and EAD. Ultimately perhaps, OAI might ally itself with the OpenURL Framework or technologies yet to be developed to broadcast researcher's queries.

When researchers dig deep, they search for the fine detail of texts and tables. Three types of text seem most essential: manuscripts and letters, newspapers and oral histories. These sources embody history in personal and immediate manner. Within a collective framework, challenges are several to the construction of research-support utilities for these materials. Text Encoding Initiative (TEI-DTD) tags are employed commonly to mark-up manuscripts, letters and oral history transcripts. But, implementations and qualifications of tags vary as widely as do implementations of qualified Dublin Core for bibliographic metadata. The same is true of newspapers. While cherry picking from the Newspaper Industry Text Format (NITF-DTD), vendors offering newspaper solutions have made proprietary modifications outside recognized standards-making processes. Whether to enable functionality or to maintain a market, these modifications make sharing and migrating text, let alone remote query of it, difficult. Scientists, particularly agriculturalists and climatologists, at the University of Florida would have us consider the numeric data trapped in text-document tables as well. Parsing a table means understanding the structure of the table.

Racing the Moon: From Diagramming the Process of Information to the Heuristics of Interfaces

Research support utilities are perhaps best first-tested among the content providers, where controls can be imposed to scientifically test hypotheses and to explore the heuristics of interface design. Doing so means diagramming research processes, having imposed upon the content providers the responsibility for understanding how research is done. Content providers, many of whom still consider themselves to be content builders, are still struggling to provide research-supporting resources. Many of us buy utilities out of the box and modify their behaviors for specific uses. At the University of Florida, for example, we knew that the aerial photograph collection was in high demand. We were familiar with the research products as well. Interviewing the collection's curators and patrons, we were able to diagram and duplicate collection uses in a digital environment.[16] Reviewing use in test and with exposure to new audiences, unanticipated uses were diagrammed and additional utilities are now being programmed. Together these dots painted a lovely picture that we quickly sold for content production dollars.

Aerial photography is an extreme example. But, the thought of exposing the collection's metadata to harvesting leads to serious questions for any collection with special, research-supporting functionality. Whether aerial photography, an herbarium specimen collection[17] or a collection of state and local history with geo-temporal tagging, today's OAI leaves functionality unexposed. It focuses upon the objects of research rather than the research itself. It has sins to eat. The researcher still has to visit each and every site of interest. If my grandmother were here today, she would probably observe that it's like attending a wake via videophone. It has the feeling of racing the moon or being told, "*you can't curl up with a computer in bed …*" Sure you can; give the technology time to catch up with the *talkies*!

Can't you just see the cow pee'n on the flatrock the *rough*? Research processes are the science fiction of OAI.

End Notes

[1] Blue and green glass bottles could capture evil spirits. A bottle-tree's primary purpose was to catch evil spirits as a Native American dream-catcher was to capture bad dreams. Today, we'll speak only of clear glass bottles and of good spirits.

[2] Research Libraries Group: Cultural Materials Initiative (http://www.rlg.org/culturalres/).

[3] University of Illinois at Urbana-Champaign's Digital Gateway to Cultural Heritage Materials (http://nergal.grainger.uiuc.edu/cgi/b/bib/bib-idx)

[4] International Children's Digital Library (http://www.icdlbooks.org/).

[5] International Children's Digital Library does not endeavor to teach as much as it endeavors to lure. It allows the child to control and manipulate the reader interface as a means of ensuring reading and to promote comprehension.

[6] E.g., Telsur Project at the Linguistics Laboratory, University of Pennsylvania (cf, http://www.ling.upenn.edu/phono_atlas/home.html)

[7] Possum pie, for those who may not understand the reference, is a very sweet nut pie. *Sweet*, as in the sweet milk used in making possum pie, is used to mean *homogenous*.

[8] Learning modules can be found in various forms of exhibit-like structures such as *The Valley of the Shadow* (http://valley.vcdh.virginia.edu/) to the field exercises of the Paleontological Research Institution's Mastodon Matrix Project (http://www.priweb.org/mastodon/matrix_directions.html) and to the traditional lesson plan structure of *Linking Florida's Natural History* (http://palmm.fcla.edu/lfnh/currmat/EdModindex.html) or of the National Park Service's *The Invention Factory: Thomas Edison's Laboratories* (http://www.cr.nps.gov/nr/twhp/wwwlps/lessons/25edison/25edison.htm).

[9] The Kentucky Virtual Library, which attempts to provide young researchers with evaluative techniques (cf, http://www.kyvl.org/html/tutorial/research/), might be taken as a model.

[10] Cf, Halbert, Martin. "The MetaScholar initiative: AmericanSouth.Org and MetaArchive.Org." *Library Hi Tech*. 21:2, pp. 182-198, particularly, pp 191-193.

[11] This technique is employed by the iArchives (http://www.iarchives.com/). Unfortunately, its web site does not provide information on the technique, which can be seen clearly in the tagged text-behind-image product. This method is especially useful in conversion of low resolution and high density resources such as newspaper microfilm. For purposes of accuracy-in-discovery, it is extremely accurate, perhaps more so than any other product currently available.

[12] PALMM Collections (http://palmm.fcla.edu/) -- Publications of Archival, Library, and Museum Materials – was founded as the collaborative digital library of the State University System of Florida. PALMM members and partners now include each of the state's public universities, several private universities, historical societies, government agencies and archives at state and local levels, and art and science museums, as well as cultural institutions in the Caribbean.

[13] Her remarks are recorded in multi-media archive of the Centre College (Danville, KY) *Southern Women Writers Conference*.

[14] Alexandria Digital Library's Gazetteer may be found at http://testbed.alexandria.ucsb.edu/gazclient/index.jsp

[15] Perseus Digital Library at Tufts University may be found at http://www.perseus.tufts.edu/

[16] *Aerial Photography: Florida* is not yet available in public release, now planned for late spring 2004. When available, it will be listed among PALMM Collections (http://palmm.fcla.edu/).

[17] Research, regardless the field, acts on all manner of objects. OAI as expressed in AmericanSouth.Org is largely a portal to textual objects. The research experience will be enriched when OAI is also able to harvest metadata expressed in Darwin Core, CIMI/CHIO standard, and other specialized metadata. OAI-PMH is enabled to harvest well beyond Dublin Core (DCMI) and Encoded Archival Description (EAD), including Darwin Core and other metadata standards (cf, http://amol.org.au/oai/files/AMOL_CIMI_OAI-PMH_Workshop_BD_Enabling_Interoperability_20020618.pdf). N.B. RLG is currently working with CIMI to harvest museum metadata (cf, http://www.cimi.org/wg/metadata/Metadata_long_desc.html)