The Metascholar Initiative: AmericanSouth.Org and MetaArchive.Org

By: Martin Halbert

Abstract:

The MetaScholar Initiative is a collaborative endeavor to explore the feasibility and utility of scholarly portal services developed in conjunction with Open Archives Initiative (OAI) metadata harvesting technologies. The MetaScholar Initiative comprises two projects, the MetaArchive and AmericanSouth projects, both funded by grants from the Andrew W. Mellon Foundation totaling $600,000. These two projects have created two metadata aggregation networks connecting some 24 libraries, archives, museums, and electronic text centers. Each network has an associated portal being created under the guidance of teams composed of scholars, librarians, archivists, and technologists. The MetaScholar Initiative is studying issues such as metadata normalization, alternative forms of scholarly communication through portals, and the process of facilitating smaller archival institutions in providing better access to their collections through the OAI Protocol for Metadata Harvesting (OAI-PMH). The MetaScholar Initiative is based at Emory University in Atlanta, Georgia.

Keywords: Data collection | Archiving

Article:

Introduction

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) enables new forms of access to the research materials held in digital repositories. The designers of the OAI-PMH carefully crafted the protocol to be a "lightweight" (easy to implement) open standard for sharing metadata that could be used to add functionality to existing and planned digital repositories. This proved to be a very effective strategy for fostering rapid adoption of the protocol.

However, the capability of repositories simply to share metadata does not by itself provide the improved access envisioned by the designers of the OAI-PMH. The designers postulated that various kinds of intermediary agencies would arise to "harvest" or aggregate metadata by means of the protocol, and then provide further value-added services that made use of the aggregated

metadata. At a minimum, such services could use the protocol to synchronize the holdings of various repositories, or might make the aggregated metadata from various repositories searchable (Lynch, 2001). The designers further hoped that the protocol might facilitate the creation of creative new services for information seekers. Such services might in fact form the basis for new varieties of online communities and scholarship.

The Andrew W. Mellon Foundation has sought to incubate and foster the creation of such intermediary services based on the OAI-PMH, and accordingly funded seven projects that advanced the concept in various ways (Waters, 2001). Two of the seven projects, AmericanSouth.Org and MetaArchive.Org, were subsequently conjoined to form the MetaScholar Initiative. This article will review the aims of the two projects, as well as accomplishments and findings to date.

Aims of the MetaScholar Initiative

The MetaScholar Initiative is a two-year endeavor funded by $600,000 from the Mellon Foundation. The goals and the rationale for combining the two original projects will be discussed in this section.

Goals of the MetaArchive project

Emory University proposed undertaking a demonstration project and feasibility study for a cross-institutional scholarly portal service providing public search functionality and subject organization for archival metadata aggregated using the OAI-PMH. The project arose from the belief that OAI-based services for researchers must do more than simply aggregate metadata. To be effective tools for research, there must be framing organization, contextual materials, and other sorts of information that add value to the basic functions of metadata aggregation and search. Further, a great advantage of such services is that they might fruitfully be focused on specific subject domains in order to concentrate and leverage attention and knowledge that subject specialists contribute to online gathering spaces. This perspective applies accumulated lessons from decades of online community research to the new opportunities provided by metadata harvesting technologies.

Case studies that informed the MetaArchive proposal included the experiences of scholars seeking to explore archival holdings related to particular topics, but distributed across many separate and often distant institutions. Such scholars often had physically to visit many separate archives to gain a comparative understanding of the primary sources of a particular subject, and were frustrated at this frequently costly and time-consuming process. This is a common situation in many disciplines that utilize the primary research materials found in archives. This appeared to be a situation that might greatly benefit from services aggregating and organizing metadata from archives with holdings in similar subject domains. MetaArchive sought to demonstrate the feasibility of addressing these needs by accomplishing four project goals during a two-year period (now at mid-point), as follows:

1. The project set out to aggregate metadata by providing partner institutions with direct assistance in the form of data conversion expertise and programming of OAI provider

systems. The project would seek to convert and import existing metadata, in the form of finding aids, catalog records, or other machine-readable forms. Typical partner institutions are archives of four-year liberal arts colleges, which frequently have only one archivist and lack technical staff or infrastructure for sharing metadata via the OAI-PMH. This category of smaller archive was a major focus because there are a very large number of such repositories that collectively hold information of great interest for scholarly research, but for which there are inadequate mechanisms for cross-institutional discovery of resources. The project also sought to work with a small number of larger archives in research universities and museums. It was anticipated that this activity of metadata aggregation would go on throughout the duration of the project, as some sites would be relatively easy and others would take more time.

2. The project seeks to establish an infrastructure serving the following purposes:

- harvesting metadata from OAI-PMH compliant sites;
- providing a public Web interface able to perform keyword and field delimited searches on the metadata records aggregated in the system;
- offering some portal functions aimed at particular scholarly communities organized around subject domains; and
- OAI data provider services (preferably installed at partner sites, but hosted by Emory if necessary) capable of exporting partner site metadata via responses to OAI-PMH queries.

The focus of much of the first half of the project was on creating this infrastructure.

3. An evaluation of the benefits of the services created will be accomplished by several means. Online surveys will be conducted through the Web interface. Additionally, focus groups of both scholars and metadata providers will be conducted to obtain feedback on the service, and suggestions for future directions. This activity will largely be undertaken in the second half of the project, once the basic service was available to evaluate.

4. An analysis of sustainability and costs of this kind of service will also be conducted, in order to better understand how other similar services might be affordably created in the future, as well as laying out a plan for mainstreaming the MetaArchive project services themselves.

Two subject domains were identified: first, papers of major political figures, and second, institutional records of religious organizations (church denominations, for example) and associated religious figures such as bishops. The subject domains were selected for the following reasons: they had interdisciplinary value as primary sources for many kinds of scholars; and they represented two coherent and representative types of archives held by many institutions that might serve as exemplars for subsequent projects. The institutional partners selected intentionally included a wide range of institutional types. Examples of MetaArchive partners include: Southwestern University, the United Methodist Archives, the Atlanta History Center, the University of the South, Davidson College, Washington & Lee University, the University of Richmond, and sub-units of Emory University and the University of Georgia.

Goals of the AmericanSouth project

The AmericanSouth project seeks to create a definitive scholarly portal for Southern history and culture. The project was the result of an extended planning effort by SOLINET. The project was similar in many ways to the MetaArchive project in that it proposed layering portal services on top of a central metadata harvester that would aggregate information from cooperating partner libraries. A team of senior scholars (rather than librarians or archivists) was to provide the intellectual organization for this scholarly portal, designing an interactive structure to promote and facilitate research, teaching, and communication.

AmericanSouth seeks to establish OAI provider systems at large research libraries situated around the Southeast. The goal is to create an extensive base of information useful for Southern cultural studies by aggregating metadata from many important archival collections held by ten institutions: Auburn University, Emory University, Louisiana State University, the University of Florida, the University of Georgia, the University of Kentucky, the Kentucky Virtual Library, the University of North Carolina at Chapel Hill, the University of Tennessee at Knoxville, and Vanderbilt University. Large research libraries were selected that had significant technical staff and infrastructures capable of installing and maintaining their own OAI provider systems, at least provided that they received a certain level of catalytic assistance by expert OAI consultants. These consultants include computer scientists from Virginia Tech who are participating in the core design work of the OAI protocol.

The project envisioned creating an online community interested in the culture and history of the American South, an active and broad subject domain. Scholars from many disciplines engaged in research and teaching are the primary audience for the service, although students and the general public are also identified audiences. The online community will provide facilities for many new forms of scholarly communication, for example subject guides, thematic articles, commentary, Web site annotations, and other contextualizing content. The project Scholarly Design Team (SDT) was to provide design guidance and experiment with these services and actively promulgate the portal in the field.

Conjunction of the two projects

The Mellon Foundation funded both the MetaArchive and AmericanSouth projects, with each project receiving approximately $300,000. Subsequent discussions between SOLINET and Emory University concerning opportunities for collaboration between the two projects led to several realizations:

- the two projects were very similar in many of their aims;
- the differences of the two projects were complementary in interesting ways, potentially enabling comparative study of several related issues; and
- market research and feedback by SOLINET member institutions indicated that the AmericanSouth project would be more successful if it was directly hosted at a major research university rather than a corporation.

SOLINET had also simultaneously lost several key personnel who were to have worked on the project.

For all these reasons, SOLINET approached Emory University with a proposal to join the two projects and operationally undertake them both at Emory as a complementary and unified initiative. Emory University accepted the proposal, and took on operational responsibility for AmericanSouth at the beginning of 2002.

The two projects quickly became closely intertwined after the decision to merge operations, and the need arose to articulate and clarify the two projects as a coherent whole to external groups. The overall collaborative effort of the two projects has been termed the MetaScholar Initiative, deriving from the prefix Meta- (for metadata) and Scholar (to reflect the fundamental orientation of both projects to the needs of scholars).

Metadata aggregation accomplishments to date

The following sections summarize the major accomplishments to date. In each case, the specific goals and outcomes will be identified.

Central harvesting and indexing system

The creation of a reliable and robust central system infrastructure to handle OAI harvesting and indexing functions was the most critical goal for the first half of the AmericanSouth and MetaArchive projects. The central system infrastructure had to be capable of routinely harvesting metadata from the many partner sites forming the proposed metadata aggregation networks of the two projects. The system also had to be capable of providing search and retrieval services for this metadata in order to make it useful to scholars. Given that the base of information anticipated in the two services was intended to be large and comprehensive, a design criterion was that the software had to be scalable to hundreds of thousands of records. Finally, the software used had to be open source software (OSS). The open source requirement was both a stipulation of the funding agency as well as a strategic commitment of the Emory libraries.

The option of project staff developing a homegrown open source software package specifically for the project was explored. An analysis conducted in late 2001 through early 2002 included consideration and testing of available open source components that might be built upon to create a harvesting and indexing system. Serious consideration was given to the possibility of building a system on the Berkeley DB toolkit[1]. This approach was abandoned for several reasons. First, it was known that several academic projects were already developing combined harvesting and indexing systems. Second, developing a software system of this complexity is a complex task that would by itself occupy a great deal of staff time. Third, by collaborating with one of the other existing academic projects, there was an opportunity to pool efforts and connect up with a larger development community. The collaborative option was ultimately selected as the best way to proceed, and an evaluation of OAI harvesting and indexing systems was undertaken.

Two systems emerged as the most promising candidates for collaboration. The first was the Open Digital Libraries (ODL) system under development at the Digital Library Research

Laboratory at Virginia Tech. This software was being developed by the research group of the project's chief technical consultant, Dr Edward Fox, and was a fully featured solution for harvesting and indexing metadata through the OAI-PMH. The second possibility seriously considered was the ARC system[2] developed by the Old Dominion University Digital Library Research Group[3]. ARC also represented a comprehensive solution to harvesting and indexing. Both research groups were approached for more information about two areas: making their software systems open source, and the possibility of collaboration with the MetaScholar Initiative. Both research groups indicated their willingness, and an extended evaluation was undertaken to determine which was the best option technically.

After evaluation, the ARC was selected for the central harvesting infrastructure, and ODL for the data providers to be set up on partner sites. ARC is an integrated OAI harvester and indexing system implemented in server-side java operating on a SQL database (for an open source implementation, either MySQL or PostgreSQL may be used). ARC only indexes the unqualified Dublin Core metadata elements served up by OAI provider systems harvested.

Both research groups have been enthusiastic about collaboration with the MetaScholar Initiative. At the urging of the MetaScholar project staff, the Old Dominion research group agreed to put the ARC software on the SourceForge site with the release of version 2.0 of the system[4]. All parties involved see many advantages to this, notably including the expansion of the programmers working on enhancing the ARC software. The MetaScholar programming team has been gaining expertise in enhancing the ARC system, and is developing add-on modules for ARC on SourceForge.

The AmericanSouth metadata aggregation network

A major differentiating feature of the MetaScholar Initiative is the collaborative cross-institutional work undertaken with partner sites to foster adoption of the OAI-PMH and actively build up a base of relevant metadata. In the case of the AmericanSouth.Org project, this took the form of creating a metadata-harvesting network of OAI provider systems installed and maintained at partner research libraries. This challenging goal entails a series of three-way collaborations between the MetaScholar project staff, partner institution staffs, and the project's OAI consultants at Virginia Tech. Each collaboration requires:

- the development of a working relationship between the three groups;
- reaching a shared understanding of the local research collections relevant to the subject domain of the project;
- reaching a shared understanding of the metadata relevant to the research collections under consideration, as well as the local systems infrastructure in which this metadata is maintained; and
- a customized solution for creating and grafting an OAI provider system on top of the local metadata infrastructure.

Each collaboration has been complex, unique, challenging and intriguing with rewarding encounters among talented people working toward a common goal.

The AmericanSouth.Org technical consultants at the Virginia Tech Digital Library Research Laboratory were responsible for the difficult task of assisting partner libraries with the creation of OAI provider systems, working to graft OAI provider scripts onto existing bodies of metadata records. The consultant team has also participated in many project meetings, and offered invaluable technical advice on many occasions on the OAI-PMH. The knotty process of setting up so many OAI providers in multiple infrastructures has proceeded surprisingly steadily, with AmericanSouth metadata provider systems in operation at four partner sites at the time of writing, including the University of Tennessee at Knoxville (UTK), Auburn University, University of North Carolina at Chapel Hill (UNC), and Emory University. Work is progressing on establishing OAI providers at the remaining AmericanSouth partner sites, and no problems are anticipated in completing the implementation of these systems within the timeframe of the project.

MetaArchive metadata conversion and aggregation activities

One aspect of the MetaArchive project that differentiates it from other OAI projects is that it has studied the issues entailed in creating a metadata aggregation network with smaller archival institutions. This is an important issue for study, because holdings of such smaller regional archives are a truly vast and "dark" scholarly Web that is a prime prospective beneficiary of the OAI-PMH technology.

In order to establish effective discovery services for aggregated metadata concerning the holdings of smaller archives, there must be catalytic nodes of support services that can help such institutions share their metadata via the OAI-PMH. The MetaArchive project has sought to illuminate the relevant issues of such catalytic support services by functioning in this role and creating a modest metadata aggregation network for focused subject domains.

The MetaArchive proposal devised a process for working with smaller archives to help them share their metadata via the OAI-PMH. The process for such data conversion activities will typically include six steps:

1. detailed interviews (usually conducted by phone, but sometimes including site visits) with personnel from contributing partner institutions to understand the structure of the metadata that they are contributing;
2. transmission of unprocessed "raw" metadata to the central project staff at Emory University;
3. identification of equivalencies between fields in the unprocessed metadata and Dublin Core elements;
4. development of PERL scripts and XSL transforms to convert and import the metadata into an OAI provider for the converted metadata;
5. confirmation with contributing institutions that the metadata has correctly been represented (accomplished by having the contributing institutions review the metadata in a staging area); and
6. movement of the metadata from the OAI provider into the harvester's publicly accessible database.

Our preference is always to install OAI data providers on partner site servers whenever possible, but if the partner site does not have the ability to install OAI data provider system locally (because of permissions restrictions or lack of technical staff time), then OAI data providers are set up on Emory servers for the partner site metadata, with provenance clearly indicated (such remotely hosted OAI data providers are sometimes called "surrogate" providers in OAI discussions).

MetaArchive has now completed many metadata conversions with project partners, including Southwestern University, the United Methodist Archives, the University of the South, and various libraries and archival units at Emory University. Many more conversions are scheduled with other partner sites. These metadata conversion activities are slated to continue throughout the remainder of the project. The metadata conversion and aggregation activities of this project have been by turns intriguing, exhausting, amusing, and very rewarding. The project staff has gained a great deal of proficiency in working with staff and institutions at many levels of expertise, size, and complexity. The project staff has also amassed a great deal of practical expertise in converting, providing, harvesting, and indexing many forms of metadata by means of the OAI-PMH. Some highlights of this experience will be summarized in the section on findings.

Creating portals and online communities

The question of what comprises a successful portal or online community for scholars has been the focus of much debate in recent years. The MetaScholar Initiative is a collaborative exploration by scholars, librarians, and archivists into the question of what kinds of online scholarly communities may be productively built on a foundation of metadata aggregation services.

AmericanSouth Scholarly Design Team

The design of the AmericanSouth.Org portal is guided by a Scholarly Design Team (SDT) of five major scholars of Southern culture and history who have been recruited to think systematically about the issues entailed in such a system. What are the requirements for an authoritative online portal for scholars? What new forms of scholarly discourse are facilitated by an online community built in close proximity to organized metadata concerning primary research materials? How can contextual functions such as annotation, interpretation, and methodological/pedagogical guides be contributed for such metadata aggregation services? In directly recruiting a team of major scholars to work on the project, AmericanSouth.Org is unlike many other such projects, which are primarily driven by technologists. The project team has consistently felt that this scholarly involvement represents a significant strength of the project.

Criteria for members of the SDT resulted from conversations with nominated SDT chair, Dr Charles Reagan Wilson, in March 2002. Members of the SDT must:

- be noted scholars in some major aspect of Southern culture and history;
- have an aptitude for grasping, exploring, and assessing the utility of online services for scholarship;

- be suited to work in a collaborative group setting such as the SDT; and
- collectively bring to bear a complementary mix of disciplinary and other perspectives.

Based on these criteria, candidates for the group were identified and recruited during the summer months of 2002. The SDT includes the following scholars: Dr Charles Reagan Wilson (Ole Miss), Dr Allen Tullos (Emory University), Dr Will Thomas (University of Virginia), Dr Lucinda MacKethan (NCSU), and Dr Carole Merritt (Herndon Home historic site, Atlanta).

The AmericanSouth SDT was convened in the latter half of 2002. The group has now engaged in the study of questions of subject boundaries, controlled vocabulary, and the nature of online publication and portals. The group has begin actively contributing content to the portal in 2003, as well as developing broader participation and knowledge of the AmericanSouth.Org site.

MetaArchive Subject Portal Working Group

In addition to the work being undertaken in the AmericanSouth SDT, the MetaArchive Subject Portal Working Group (SPWG) has considered the question of how aggregated metadata can be made coherent and useful to scholars. The group comprises librarians and archivists with doctoral level subject qualifications and extensive experience in the field.

To gain a better understanding of the needs of users, the SPWG and MetaScholar development team periodically conduct focus groups with researchers, faculty and staff members, students, archivist and librarians from various institutions. In each session, attendees are typically given background information on the project, a demonstration of the current version of the system, and then engaged in structured feedback discussions. Five to seven individuals usually participate in each group.

The results of focus groups to date have provided helpful feedback and frequently been quite surprising. Virtually all participants consistently feel the service as demonstrated would be helpful for scholarly research, but many express a healthy caution concerning new online services and ask probing questions concerning how the service will ensure that the metadata will be accurate and trustworthy. Participants have been much more technically astute in the use of online resources than had been previously assumed, and often ask for many additional complex searching features.

Focus group participants have expressed interest in various kinds of contextualizing features, but different categories of researchers had quite different expectations for the portal service. Beginning researchers were found to be much more interested in contextualizing guides to methodology and pedagogy of using the primary sources the service indexed. Senior researchers were much more resistant to including such contextualizing features, being skeptical of all preconceived interpretations of information.

The feedback from the focus groups guided much of the early design thinking of the project team. Many of the features desired were relatively easy to add, for example highlighted search terms in result sets. Others were more difficult to imagine incorporating, given the nature of the metadata being harvested, for example, the ability to limit the search to a given time period.

Since there is not consistent encoding of temporal information in metadata coming from disparate sources, this capability would be hard to develop.

Software for scholarly portals and online communities

The AmericanSouth and MetaArchive projects were designed to incorporate systems for the facilitation of many scholarly activities, such as commentary, annotation, interpretation, analysis, and discourse concerning research collections described by the metadata aggregated from the underlying harvesting networks. The software enabling such activities had to be flexible, extensible, and open source in availability.

In analyzing these requirements, the project staff decided very quickly, that it was again not reasonable to try to develop such software from scratch, and instead began to evaluate the many new open source software projects now emerging to address broad system functions variously described as portals (an enormously contested term and topic) and content management systems[5].

Systematically evaluating these systems depended on many factors difficult to determine conclusively at the beginning of the project, such as what features the AmericanSouth project SDT would eventually settle on as desirable, and which capabilities the MetaArchive SPWG would identify as critical. In spite of these uncertainties, the project staff felt that they had to select some tool in order to begin developing an initial user interface to the AmericanSouth and MetaArchive portals. This was seen as essential because both the SDT and SPWG needed some system to respond to conceptually in order to understand basic concepts of portal design and possibilities. The tool needed to be simple to use, be modular and adaptable, and have enough features that it would rapidly jumpstart the process of brainstorming new approaches to scholarly communication with the SDT and SPWG. Many systems were evaluated, and experiments were undertaken with several packages, including Zope[6] and OpenCMS[7].

The tool eventually selected was PostNuke[8], a popular open source software system with portal and some content management features. The software has an open architecture based on common OSS tools such as the PHP programming language and MySQL. PostNuke has an active development community coming from many different countries and fields. Because of its open and modular architecture, the project staff felt confident that it could be easily adapted to any of the various OAI harvesting and indexing systems that were simultaneously being evaluated (this indeed proved to be the case).

The large numbers of PostNuke modules that have been developed and continue to appear mean that the project has many ready-made options for new communication tools to present to the SDT and SPWG for evaluation. Examples of these include Web site annotation tools, threaded commentary and topical discussion forums. These tools all represent best practices from years of online community experiments. These tools can also be adapted to facilitate activities such as scholarly interpretation and analysis. The project staff has already had success in developing ad hoc PostNuke modules, and feel that if the SDT or SPWG expresses a desire for some particular new functionality, it can be developed relatively easily and rapidly in the PHP environment of PostNuke.

Downsides of the PostNuke software include some concern about security features in the system, and a general anxiety over whether or not it will be appealing and suitable to the SDT and scholars in general. However, since there was no practical way to establish the superiority of any particular portal software prior to convening the SDT and SPWG and conducting usability testing, the choice of PostNuke seems as good as any other system the projects might have adopted. There is also nothing to preclude changing to another system should PostNuke prove unsatisfactory, given the flexible nature of the underlying PostNuke database and its export capabilities.

Selection and discovery of additional scholarly resources

The AmericanSouth and MetaArchive project proposals described aggregating metadata concerning an assortment of particular scholarly resources at particular research institutions. In addition to these specified research collections, an expectation was articulated that additional relevant collections and material would be opportunistically identified during the course of the projects as relevant for inclusion. These opportunities might occur systematically by project staff studying the partner institutions, through discourse among scholars, during site visits to the partner institutions, or in other unplanned ways during project collaboration activities.

These opportunities have indeed arisen. As project staff members have collaborated with partner site staff, additional research collections have been identified at AmericanSouth and MetaArchive partner sites. Some examples include Auburn's Transforming America collection on Black history and civil rights in Alabama[9], Emory's Southern Changes Online collection from the Southern Regional Council[10], and several of the new digital library collections at the University of Tennessee at Knoxville[11]. Most of these collections did not exist at the time that the grant proposals were submitted, and project staff heard about them while they were under development during meetings with the partner institutions.

As AmericanSouth.Org was conjoined with MetaArchive.Org, many more opportunities for additional research collections arose. The vast majority of the collections identified from the beginning for inclusion in MetaArchive are natural additions to the AmericanSouth subject domain, since they primarily represent archival collections of Southern political figures and religious institutions. More than 24 collections from various MetaArchive sites concern Southern topics, and will in all likelihood be included in AmericanSouth.Org. As project staff entered into discussions with MetaArchive partner institutions, they also began to uncover previously unknown collections that were of relevance to AmericanSouth subject domains.

All of these opportunistic additions were desirable occurrences, but they had the effect of further complicating the process of collaboration. This was because collection evaluation activities had to now be factored into the work, as well as increasing the number of custom OAI provider systems that had to be developed for AmericanSouth and MetaArchive sites.

Interesting boundary classification decisions arose in these evaluative activities, in a way similar to collection development activities in libraries. Were particular collections relevant to AmericanSouth? MetaArchive? Both? The AmericanSouth project proposal anticipated the

development of criteria for selecting resources, and indeed this was necessary. However, these criteria decisions were sometimes hard to adjudicate. An example was the Martyred President Sermons, an online collection of 57 sermons given on the occasion of the assassination of Abraham Lincoln[12]. The sermons are concerned with Lincoln's assassination, a topic of relevance to the Civil War and potentially AmericanSouth, but all the sermons were delivered in northern states. The documents concern a major political figure (a MetaArchive subject domain), but are not from archives of Lincoln's papers. The sermons are religious institutional documents, another MetaArchive subject domain, but very different from other items in the MetaArchive collections. The metadata for the sermons will be harvested for AmericanSouth.Org, but all such harvesting decisions will be reviewed subsequently by the SDT.

Another category of relevant metadata that became obvious in the first months of the project work were research collections for which OAI provider systems had already been established, notably American Memory. The complicating issue is that of harvesting only relevant records from a body of metadata as broad as American Memory, which does not implement OAI sets in such a way that records associated with the AmericanSouth can be harvested. A keyword match approach is now being investigated in conjunction with input from the SDT. The project team intends to harvest records from other OAI provider systems such as American Memory, once an acceptable filtering process can be developed.

Initial findings

The MetaScholar Initiative has explored many aspects of OAI metadata harvesting services as they relate to scholarly communication. Some of the findings of the project to date are details of issues that were anticipated, but many are unanticipated surprises. In this section, some of the more interesting and unexpected findings are reviewed.

The OAI-PMH: successes and challenges

The OAI-PMH is the core research focus of all the projects that collectively comprise the Metadata Harvesting Initiative of the Mellon Foundation, and it has been the target of intense study in the MetaScholar Initiative.

First of all, it must be said that the MetaScholar projects can confirm unequivocally that the protocol works effectively as a mechanism enabling automated metadata distribution nodes and harvesting networks. This is by now an unsurprising finding, as many projects have also confirmed this and have been using the protocol in the construction and planned construction of digital library systems that share and aggregate metadata. But such confirmations are worth explicitly stating. The OAI-PMH works well. The concept and architecture are sound and effectual.

Second, the MetaScholar Initiative has now undertaken many collaborative efforts with partner institutions of various sizes and types, and can begin to report on patterns that emerge in grafting OAI functionality onto existing metadata infrastructures. A frequent scenario that we have encountered in both the MetaArchive and AmericanSouth projects is that of adding PERL scripts to an existing directory tree of metadata, expressed as either XML files or HTML that can

dynamically be converted to XML. Virginia Tech has developed a variety of such scripts that can be deployed with minimal effort[13].

OAI functionality can also be added with very modest effort to repositories that store their metadata in some kind of database management system. The typical DBMS now in use by libraries and archives is accessed via some sort of Web API (Application Program Interface), often through a programming language such as PHP or PERL, and adding the six OAI operations to such an API is a straightforward task.

The final category of infrastructure we have worked with is that of the integrated library system (ILS) or online catalog. While relatively few archives store their metadata solely in their institutional ILS, this is an interesting and complicated enough situation that a separate section is devoted to it below.

Challenges of collaboration for metadata harvesting

Collaboration is difficult

Collaboration among many institutions is exceedingly difficult. In the process of working to coordinate the collaboration needed between AmericanSouth and MetaArchive participants, the project staff have had an opportunity to consider many issues that arise during such group efforts. What are the motivations that obviously may cause separate agencies to work together? What are the unobvious motivations that may be appealed to, and drawn out in discussions? What are the root circumstances that arise to stymie collaboration?

Most academic libraries and archives today feel chronically under-funded and under-staffed. The first obstacle to collaboration is simply that libraries may in practice feel wholly incapable of devoting significant effort to any new externally oriented project, whatever the potential benefits. Indeed, many feel incapable of adequately handling their own mission-critical responsibilities. This is especially true for the smaller MetaArchive institutions, but is also the case for many of the large AmericanSouth research libraries. In this situation, even if the administration and staff at such institutions are very interested in the project, it becomes difficult to engage them in a sustained way, especially if they have lost digital library staff who held essential knowledge and expertise regarding local infrastructures. In the extreme case, it is obviously not possible to set up an OAI provider for a library if no one there any longer remembers where or how their local metadata records are maintained. Also obvious is that the library in such a situation has more pressing problems than collaboration. The first obstacle to collaboration on digital library initiatives is therefore simply the problem that many research libraries lack the staff to engage effectively in the work.

Another problem may be that local digital library metadata systems are in transition. This has also proven to be the case in several AmericanSouth partner institutions. Previous first generation or experimental systems are being migrated to new systems. It is evident in this situation that an OAI provider should not be designed for a system that is being phased out and likely to be retired within the immediate future. But if the new infrastructure is not yet operational, or even clearly understood as yet by the partner library staff, one simply has to wait

for the local situation to stabilize before progress can be made toward OAI services to be layered on top of this local infrastructure.

Metadata quality is sometimes an issue that has arisen. This is surprising because the partner institutions involved in the project are all research libraries, and such institutions should presumably produce high-quality metadata. Yet we have struggled to respond appropriately to (a very few) situations in which unacceptable metadata gets heaved up from newly installed local OAI provider systems, for example, including many records without so much as a title in the Dublin Core elements. The reason for this was most often the fact that the metadata in question was still being collated from source systems, and was moved into OAI data providers in too early a stage of processing. We have worked with our partners collaboratively to identify and remediate all cases of inadequate metadata noted.

The need to improve understanding of the OAI-PMH

As mentioned, there is still a widespread lack of understanding about the OAI-PMH and the uses for which it is intended. There is a long way to go until the protocol is internalized and accepted to the same degree as MARC. How big a problem is this? Is it just a matter of waiting for people to eventually "get it", or are there dependencies that have to be addressed here?

One of the things that the author has regularly done at conference presentations is to poll the audience with three questions:

1. Question number one: "How many of you were familiar with the OAI before this presentation?" Answers have ranged between 50-75 per cent.

2. Question number two: "How many of you are from institutions which currently operate OAI provider systems?" Answers have ranged between none to 15 per cent, frequently with indications that many people have no clue about what their institutions do concerning OAI.

3. Question number three: "How many of you are from institutions which are currently planning to establish OAI provider systems?" Answers have ranged between 30-70 per cent, sometimes with hints that people think their organization will set up OAI systems simply because they have become convinced at the presentation that OAI is inherently a good way to go, or perhaps just an inevitable idea.

The fear that arises is that OAI is like dieting or other well-intentioned practices: something most people know about, most people are planning to do some day, but which few people ever effectively do in practice. There appears therefore to be a need to improve understanding of the OAI-PMH in the field of librarianship and archives in order to ensure that the protocol fully realizes the potential benefits it was designed to produce. A few ideas about fostering the adoption of the protocol will be laid out in the section on next steps.

Challenges of metadata format collisions

One unanticipated sort of problem that occurs in practice when creating OAI provider systems is what we term conceptual collisions of metadata formats. The OAI-PMH is flexible in that it allows OAI provider systems to serve up any metadata schema that can be validated against an available XML Schema Definition, though all OAI providers are prescriptively forced to serve up metadata in at least unqualified Dublin Core (UDC).

When a project is made responsible for the creation of OAI provider systems on behalf of various sub-agencies or entirely separate institutions (as have several of the Mellon-funded projects, including at least the two MetaScholar projects and the University of Illinois project), the project staff are faced with many practical decisions about mapping metadata from one representation into UDC. It quickly becomes obvious that there are some metadata formats that map (or "crosswalk") into one another better than others. A prime example of this problem (although not the only one, by any means) is the case of the Encoded Archival Description (EAD) and UDC. EAD is a strongly hierarchical format, reflecting the practice of archivists in describing a collection in a single relatively complex document, rather than in a large number of granular entries ala a card catalog or typical database.

What this means in practice is that it is hard to represent a finding aid encoded in EAD in UDC records. The programmer (or other individual developing a crosswalk) will almost inevitably be forced to degrade the metadata using one of two general strategies, both of which are classically Procrustean approaches. The first option is to compress the complex EAD record into a much more primitive and terse single UDC record. This approach inevitably discards much information in the form of series and item level information, at least if the resulting entry is to be at all comparable to most UDC records. The second option is to chop up the EAD record into a myriad UDC records that are even more clipped, but which together capture all the detail of box and folder level data in the finding aid. The obvious difficulty of this strategy is relating the individual UDC records that may thus be encountered in the system to the whole collection. Box and folder information in finding aids is notoriously brief and idiosyncratically expressed. How useful is it to encounter a record with the cryptic title "Box 456 – Letters to Burt – 1956"? There are equivalent, but separate issues that arise for other metadata formats and UDC.

This is a very thorny problem, and it has received a great deal of thought in recent months as MetaScholar has encountered the issue repeatedly. The University of Illinois metadata project has also studied the question, and presented their misgivings and ad hoc suggested solutions (Habing and Prom, 2002).

Some of the reasons the problem is so troubling are because one is forced to degrade the metadata no matter what strategy one takes, there are no obvious approaches to resolving the issue or making progress on it that will make anyone happy, and until this problem is settled in a predictable way, it is difficult to make reliable quantitative estimates of the number of records that are likely to be aggregated in the system over time. The MetaScholar project staff and officials of the Mellon Foundation believe that the overall goal should not be to simply maximize the quantity of metadata records aggregated in the system, but rather to maximize the utility and quality of the records that are produced and aggregated in the process. MetaScholar is soliciting input from archivists and others during focus groups in studying this question.

Other metadata quality issues

There are many additional unexpected theoretical issues that arise in metadata aggregation activities. Some particular discussions, summarized here, that the SPWG has engaged in that led to surprising conclusions include authority control and de-duplication strategies. The MetaArchive programming team has had many long conversations with SPWG members Susan Bailey (Emory Division Leader for Bibliographic Gateway Services) and Susan Pinckard (Emory Monographic Cataloging Team Leader), and the project staff is grateful to both for their patient guidance and expert advice in these matters.

Authority control

Authority control is one of the most critical strategies ever developed for managing metadata in the many years of practice in maintaining library catalogs and other bibliographic databases. Yet, authority control is virtually impossible in the rough-and-ready OAI world of heterogeneous metadata streams, where the only common ground is the loose standard of unqualified Dublin Core metadata elements. What the project staff realized early on is that we had actually had plenty of authority control in the incoming metadata from our providers, too much in fact! This is because each partner had their own respective authority control mechanism, virtually none of which were held in common. Southwestern had an internally developed set of subject headings for the Tower papers, Sewanee had various controlled headings for the Episcopal material, etc.

All of these were logical approaches in the context of the microcosm of the individual archives; it only becomes an issue when all the metadata is aggregated. But the point of authority control is to enable consistent discovery of related items. The project staff began to wonder if this was actually a non-issue, since the contents of all the different archives were so different that the concept of browsing closely related items only made sense in the context of a particular archive. One of the purposes of the overall service is of course to discover related items cross-institutionally, but it may be that this can only happen in practice through keyword searches. We have decided to take the general strategy of preserving authority control wherever it occurred in incoming metadata streams for discovery of related items within an archive, but to abandon the search for a mechanism for adding authority control to the entire database.

De-duplication strategies

The point of de-duplication is to avoid multiple "bad" hits in retrieval sets that are "bad" because they all point to the same content, and are therefore irrelevantly repetitive. Duplicated records may be likely if a metadata aggregation service is harvesting from databases that may include records for the same items, primarily published material that may be held by multiple institutions.

If the service is harvesting metadata concerning unique items (such as archival one-of-a-kind manuscripts), this is less of a danger. "Traditional" metadata management systems like online catalogs may attempt to detect duplicate records after the fact of introduction by matching on unique identifiers like ISBN numbers, or (less reliably) on titles or other bibliographic data. In

the OAI gallimaufry of different kinds of "stuff", there are usually no unique identifiers to match on.

The MetaScholar approach has been to try to de-duplicate metadata before the fact of introduction by limiting metadata to unique archival items, and having a close understanding of what items are in the different OAI providers we harvest from. Since project staff members have worked closely with all our partners on creation of their OAI providers, this has not been a problem for us to date. But this will likely become a problem for us in the future if we proceed with some of our thoughts concerning the inclusion of broader item formats, notably books (see further discussion below). In general, we tend to agree with those who maintain that duplication of records may be a frequent downside to the flexibility of the OAI protocol, but that this is more of an annoyance than a showstopper to metadata aggregation services.

Legacy print metadata

As the project team conducted initial interviews with archives to evaluate which collection metadata would be of value for inclusion in the subject domains of MetaArchive, we often had a particular sort of unpleasant experience. After identifying and discussing at length a wonderful newly discovered collection with rich item level records held by the partner institution, we then ran into a brick wall when we discovered that all of the metadata for the collection was in the form of printed catalogs. Since recon efforts were explicitly outside the scope of the project for many unavoidable reasons, this was a very frustrating sort of occurrence. Project staff members sometimes call this the "stone tablet" syndrome; an archive might have God's own scrapbook, but if it is carved on stone tablets, you are looking at recon work.

The general observation here is that to form a comprehensive subject domain portal, one must consider where the relevant metadata resides. The OAI-PMH is a logical mechanism for sharing metadata that already resides in a machine-readable form, and whose maintainers are willing and able to distribute their metadata via creation of an OAI provider. There may be bodies of relevant materials whose metadata are not digitized, and whose maintainers are (for any of many different reasons) unable or unwilling to bring up an OAI provider. For such bodies of metadata, an aggregation service will need to first undertake other enabling activities, like recon or other forms of metadata aggregation.

Surrogate OAI provider issues

Metadata aggregation networks like MetaArchive comprise a distributed, very loosely coupled confederation of institutions that have widely varying levels of technical infrastructure. Many institutions with important research materials may lack the infrastructure to serve out their metadata via OAI. A key point of the MetaArchive model is that a central catalytic institution can serve as a nexus of metadata facilitation services, either providing expertise to advise smaller yet capable institutions in setting up an OAI provider, or in outright migration of the smaller "client" institution's metadata into a metadata provider maintained at the larger catalytic "host" institution. This latter option may be termed a surrogate metadata provider, because the larger institution is acting on behalf of the smaller institution to share its metadata with the community.

The concept of surrogate metadata providers raises issues, however. All the respective roles and responsibilities of the client and host in the surrogate metadata provider model are not clear from the outset, as this concept is new and untried. The MetaScholar project team has raised concerns during planning meetings relating to this point. It seems logical to develop a memorandum of understanding that clarifies such expectations between host and client institutions. The MetaScholar project staff is discussing the issue with the Mellon Foundation to identify what language such a memorandum should include. This may subsequently serve as at least one example of how to manage expectations around roles and responsibilities in surrogate metadata provider relationships.

The OPAC and the OAI

Roughly half of the MetaScholar Initiative's partner sites store some of their metadata records in their online catalogs. Different libraries adopt many different views of what types of archival information should be included in their online catalogs. At some institutions the online catalog is seen as the primary discovery and metadata maintenance system. An example of this viewpoint is LSU, which records the vast majority of special collections information in the online catalog. The opposite extreme of all archival records in separate databases might be represented by Southwestern, where the special collections Procite database records all archival holdings. The position in the middle might be represented by UNC Chapel Hill, which hedges bets by recording at least all collection level records in both their OPAC as well as their special collections Web systems. All of these strategies are viable and consistent with good archival practice.

This squarely raises the question of the relationship of the OAI-PMH and the "traditional" library metadata repository of recent years, namely the online catalog. MetaScholar staff members have come to believe that there is an important set of linkages between the OPAC and OAI systems, both providers and harvesters.

The OAI-PMH is now being adapted and added on to online catalogs just like it has been added to so many other systems. In support of the Unicat Virtual Union Catalog of Belgium project, Benoit Pauwels has developed a package of PERL scripts that add OAI functionality to the SIRSI Unicorn ILS[14]. The MetaScholar project team is working with Emory's ILS support group as well as the LOUIS consortium in Louisiana to test OAI services on top of several SIRSI Unicorn sites using the Belgian scripts. Adding OAI provision services on top of online catalogs raises many questions, especially around the implementation of sets in such a system. The Belgian effort was undertaken simply to enable wholesale transfer of bibliographic records from individual ILS to the proposed Unicat virtual union catalog system on a regular basis. There is no support for subject-based OAI sets; an issue that becomes relevant if one wants to extract metadata pertaining to particular topics from online catalogs. SIRSI Unicorn is undoubtedly not the only ILS for which OAI modules have been developed; it is simply an example of particular interest to the MetaScholar Initiative.

Another mechanism for quickly adding OAI capabilities to an ILS is through Z39.50, if the ILS has that capability. The ZMARCO package developed by the University of Illinois is an elegant mechanism for enabling this approach, although it imposes some potentially stringent

requirements for the underlying Z39.50 gateway in order to work. MetaScholar is exploring use of this tool with several of our partner sites.

Many archives focused on particular subject domains include rare, but not unique, books germane to the subject. Such books may be important items to take into account for scholarly portals like MetaArchive or AmericanSouth. The MetaScholar project staff has repeatedly run up against this issue in evaluating the collections of our partner institutions.

We are unsure whether or not to include bibliographic metadata for such rare book collections, as it obviously raises a hornet's nest of additional issues to deal with (de-duplication is not even the biggest issue). Where should the line be drawn? What constitutes a relevant rare book that should be included? This also relates to even broader questions about the relationships between various kinds of discovery services.

Next steps

Both AmericanSouth and MetaArchive have made steady progress toward all their project goals. Several activities will receive attention in the second half of the two projects.

Completion of the remaining OAI provider systems

The basic underlying OAI harvesting and indexing infrastructure of the MetaArchive and AmericanSouth metadata harvesting networks is now essentially complete, by means of the ARC software re-implemented by the MetaScholar Initiative staff. The primary tasks remaining in the area of technical OAI harvesting work are the completions of all remaining OAI provider systems at partner institutions, primarily based on ODL software tools. The final version of these provider systems could not be constructed prior to the finalization of the OAI-PMH version 2.0, but can proceed now that the production version of the protocol has been finalized. The development of a standardized memorandum of understanding for surrogate metadata provider relationships will be explored. The notes kept by MetaScholar project staff during these collaborations will be synthesized in the second half of the project into a set of suggested guidelines for institutions and personnel considering either the creation of new repositories of metadata or the addition of OAI services on top of existing repositories.

Further infrastructure development

Further development of additional features of the MetaScholar system infrastructure will be guided by feedback from the MetaArchive Subject Portal Working Group, the AmericanSouth Scholarly Design Team, and partner site staff. We anticipate developing additional components that will address specific activities in a modular way. For example, an idea that has been floated several times is that of a software component to enable either archivists or scholars to enter new collection level UDC records quickly and submit them into a provider and thereby to the harvester. Additional portal modules will undoubtedly be suggested by the SPWG and SDT. These will be developed, tested and evaluated as time allows. Finally, better overall documentation for the system components will be developed. This will support future

maintenance of the system, and enable others to replicate any portions of the system desired elsewhere.

Assessment and follow-up activities

There are a variety of assessment activities laid out in the project plan. Tests of OAI protocol compliance in OAI provider systems are part of the everyday work of setting up new providers and are ongoing. Additional focus groups are planned to conduct both usability assessments as well as assessments of the services offered in AmericanSouth and MetaArchive. Online surveys will be conducted of both partner library staff and users of the two portals. The SDT will be debriefed at the end of their activities concerning AmericanSouth.Org, and will act as promulgators of the service in professional presentations subsequently. Costs of different aspects of the work are being tracked during the course of the project by the MetaScholar staff. The assessment activities will guide plans for the long-term sustainability of the services established during the project.

Activism for Open Archives Initiative and Open Source

As a more general and open-ended next step, the MetaScholar Initiative will continue to advocate that software projects in the digital library arena be made open source. Two successful examples of this advocacy so far were the agreements we brokered with our technical collaborators at Virginia Tech and Old Dominion to make their OAI software packages open source. Each group had been considering the option, but did not actually take steps until pressed by the MetaScholar Initiative. All software additions to either of these systems subsequently developed by the MetaScholar Initiative will of course also be made freely available via open source license.

The MetaScholar Initiative intends to look for further opportunities to advocate and advance both the OAI and the open source movement in coming months. We will primarily do this by actively exploring possible alliances with other OAI projects and metadata discovery services.

Refinement of subject portals

We want to ensure that the subject domains are clearly articulated as coherent topics. For example, we have considered refining the articulation of the subject domain to characterizations like "Major Figures in Southern Politics and Government", or even just "Southern Politicians". The general sort of question we are studying for reference by subsequent projects is: "What are the characteristics of subject domains that are broad enough to be relevant to a cross-section of scholars, but yet targeted enough for a reasonably small and agile subject portal project?" We also want to clearly understand what approaches to loosely controlled vocabularies will be productive in organizing metadata and contextual material. For example, in the AmericanSouth project the original thinking was to structure all metadata according to a strict framework taken from the book Origins of the New South, 1877-1913 by C. Vann Woodward. The Scholarly Design Team has since reconsidered this position, and decided to adopt a much broader set of controlled vocabularies from both the Encyclopedia of Southern Culture and the Library of

Congress subject headings. These issues will be taken up by both the SPWG and SDT in coming months.

The community interaction features of the portals will also be studied intensively in the second half of the project. We intend to experiment with many suggested features from the focus groups we have held to date. Some examples of features to try out are notification services for new additions to the database, the entire threaded commentary capability of PostNuke, saved citation lists (e.g. the University of Illinois OAI search interface's bookbag feature), and ratings systems (e.g. the ratings of reviews in Amazon.Com). The exact benefits of many of these features are yet to be determined in the context of subject portals like those of the MetaScholar Initiative. This relates strongly to the question of evaluating the benefits of the service to scholars.

Evaluation of benefits for scholars

Online surveys will be designed to gauge what perceived benefits the service offers to scholars. These surveys will take two forms:

1. very brief one question surveys that will be periodically placed on the MetaArchive.Org homepage using the survey module of PostNuke; and

2. a more extended online survey linked from the homepage to either SurveyMonkey.Com or another online survey system with good capabilities for tabulating results

The design of these surveys will require careful thought as to questions that will shed light on the issue of benefits to scholars, as well as strategies for garnering reasonable numbers of responses.

Focus groups of scholars will be held to gain information concerning several topics:

- metadata normalization strategies that are preferable for scholarly purposes; and
- benefits and usability of the service.

Concerning the latter topic, we are exploring ways to do usability testing with the resources available to us. While true usability testing requires sophisticated and specialized facilities and experts, we may be able to perform some useful usability assessment with less that this scale of investment. In any event, the less expensive technique of focus groups has proved very informative and helpful, and will be used on several occasions in the last half of the project.

Focus groups of archivists and librarians, both from MetaScholar partner sites and control groups will also be conducted to study the perceived value of the service to custodians of metadata and information providers.

Sustainability study

An assessment will be prepared of the initial and ongoing costs of the MetaScholar services, as well as potential means of sustaining them. This sustainability study will attempt to shed light on the general question of what it costs to implement and operate a scholarly portal and metadata

aggregation service for a particular subject domain. The answers to this question will hopefully provide valuable information for future efforts aiming at the creation of similar services. An attempt will be made to identify the total cost of the operation, including salaries of all contributing personnel, equipment costs, and other necessary expenses of continuing to carry forward the service.

Notes

1. www.sleepycat.com (UC Berkeley SleepyCat Software, makers of Berkeley DB).
2. http://arc.cs.odu.edu (ARC, OAI federated search service software).
3. http://dlib.cs.odu.edu (Digital Library Group, Old Dominion University).
4. http://sourceforge.net/projects/oaiarc/ (Open Source distribution of ARC).
5. www.oscom.org/Conferences/Zurich/ (1st Open Source Content Management System Conference).
6. www.zope.org (Zope.org).
7. www.opencms.org (OpenCms Project).
8. www.postnuke.com (Postnuke).
9. http://diglib.lib.auburn.edu/transam (Transforming America, Auburn University).
10. http://chaucer.library.emory.edu/schanges/southernchanges.html (Southern Changes Online, Emory University).
11. http://diglib.lib.utk.edu/dlc (Digital Library Collections at University of Tennessee at Knoxville).
12. http://chaucer.library.emory.edu/lincoln/html/ (The Martyred President Sermons, Emory University).
13. www.dlib.vt.edu/projects/OAI (OAI Projects, Digital Library Research Laboratory, Virginia Tech).
14. www.kbr.be/unicat/doc/index.htm (Unicat Virtual Union Catalog, Belgium).

References

Habing, T.G. and Prom, C.J. (2002), "Using the Open Archives Initiative protocols with EAD", Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries, 14-18 July 2002, Portland, Oregon, Association for Computing Machinery, New York, NY, pp. 171-80.

Lynch, C.A. (2001), "Metadata harvesting and the Open Archives Initiative", ARL Bimonthly Report 217, August, available at: www.arl.org/newsltr/217/mhp. html (accessed 17 February 2003).

Waters, D. (2001), "The Metadata harvesting initiative of the Mellon Foundation", ARL Bimonthly Report 217, August, available at: www.arl.org/newsltr/217/waters.html (accessed 17 February 2003).